BEYOND SPATIAL CORRESPONDENCE: OPTICAL-DENSITY-GUIDED MARGINAL AND RELATIONAL LEARNING FOR WEAKLY-PAIRED H&E-TO-IHC VIRTUAL STAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly paired H&E-to-IHC virtual staining learns from adjacent tissue sections, where deformation, tissue loss, and expression heterogeneity weaken local correspondence. Spatial co-location and visual similarity do not necessarily imply compatible staining states, making them insufficient on their own for organizing stain-consistent supervision. Optical density (OD), widely used to quantify immunohistochemical (IHC) staining intensity, provides complementary evidence of stain expression. We propose PGLStain, which uses continuous OD compatibility not merely as an intensity-matching target, but as a signal for conditioning feature relations and allocating local supervision. Protein-Expression-Aware Correspondence Calibration (PECC) combines feature-context and OD-calibrated graphs, assigns relative weights to spatially proposed candidate pairs, and aligns permutation-invariant OD-state graph summaries. To complement candidate-based supervision with image-level expression constraints, Marginal–Relational Stain Alignment (MRSA) aligns unordered OD marginals and uses unbalanced Gromov–Wasserstein transport to compare intra-image expression-state relations under relaxed node-mass constraints, without fixed cross-section node assignments. Together, these objectives integrate stain-calibrated local supervision with correspondence-free expression-state learning, without additional inference-time computation. Experiments on public H&E-to-IHC benchmarks demonstrate competitive perceptual quality and quantitative staining agreement. Code are avaliable at anonymous repository

1 INTRODUCTION

Virtual immunohistochemical (IHC) staining aims to synthesize IHC images from routine hematoxylin and eosin (H&E) images, offering a computational approach to estimating protein-expression patterns from tissue morphology. In weakly paired datasets, H&E and IHC images are acquired from adjacent tissue sections. They retain partial spatial correspondence but differ locally because of deformation, tissue loss, and expression heterogeneity (Peng et al., 2026). Learning must therefore exploit the reference information without treating spatial co-location as verified correspondence.

Existing methods address weak pairing through contrastive learning and adaptive weighting (Park et al., 2020; Li et al., 2023; Zhang et al., 2024; Wang et al., 2024), feature decomposition or auxiliary-task alignment (Wang et al., 2025; Xie et al., 2026), graph-based contextual modeling (Qiu et al., 2026; Liu et al., 2026), and transport-based matching (Guan et al., 2025; Peng et al., 2026). These approaches organize supervision through candidate weighting, contextual aggregation, or soft cross-image associations. This setting raises two complementary challenges.

First, local expression correspondence is ambiguous (Figure 1(a)). Spatially nearby or morphologically similar regions in adjacent sections may exhibit different staining states. For example, ER/PR staining can vary within a single tumor (Allott et al., 2016), and regions with similar nuclear organization may differ in staining intensity or DAB-positive fraction. Consequently, spatial and visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

evidence alone may be insufficient to determine whether a local reference region provides reliable supervision for expression prediction.

Second, regional expression heterogeneity is difficult to constrain across imperfectly corresponding sections (Figure 1(b)). Similar image-wide staining statistics can coexist with different distributions of regional mean expression. Meanwhile, the locations and relative abundance of expression states may vary between adjacent sections. Agreement in overall staining therefore does not by itself establish consistent regional expression, while fixed-location comparisons may confound prediction errors with genuine section-to-section differences.

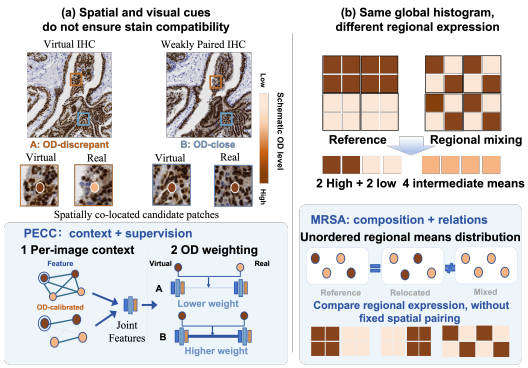

Figure 1: Spatial and visual similarity need not imply stain compatibility, and matching global statistics can miss regional expression differences. PGLStain combines OD-calibrated local supervision with regional expression alignment without fixed cross-section assignments.

Optical density (OD), a stain-intensity proxy widely used in quantitative IHC analysis, provides complementary stain-specific evidence. Prior methods exploit OD and its focal variant (FOD) through statistical and semantic constraints (Chen et al., 2024; Peng et al., 2026) or graph-node attributes (Qiu et al., 2026). Regions within the same positivity category can still differ substantially in intensity; continuous OD values retain these within-category differences. Building on this evidence, we investigate how OD can guide both local supervision and regional expression-state alignment, without treating OD agreement as correspondence ground truth.

We propose PGLStain, which addresses these challenges through stain-calibrated local supervision and correspondence-free expression-state alignment.

Protein-Expression-Aware Correspondence Calibration (PECC) addresses the first challenge through two complementary decisions: how each sampled patch is represented through intra-image context, and how strongly each cross-image candidate contributes to supervision. A feature-context graph and its OD-calibrated counterpart jointly encode each patch, retaining broad visual context alongside stain-conditioned relations. Continuous OD weighting calibrates the relative contributions of spatially proposed virtual–reference pairs to bidirectional contrastive learning. PECC additionally aligns permutation-invariant summaries of an independent OD-state graph, without index-wise cross-section matching.

Marginal–Relational Stain Alignment (MRSA) addresses the second challenge using independently extracted tissue-block sets. It aligns unordered distributions of block-mean FOD, retaining regional expression summaries without fixing cross-section block assignments. It further compares pairwise relations among regional states described by mean FOD and soft-positive pixel fraction. Unbalanced Gromov–Wasserstein transport estimates soft couplings under relaxed node-mass constraints, accommodating differences in state proportions. Together, these objectives constrain expression composition and state relations within each input image rather than requiring spatially corresponding reference regions.

All proposed objectives operate only during training and introduce no additional inference-time computation. Experiments on public H&E-to-IHC benchmarks demonstrate competitive perceptual quality and quantitative staining agreement. Our main contributions are summarized as follows:

- We formulate continuous OD compatibility as a shared signal for conditioning contextual relations and allocating local candidate supervision under weak pairing.
- We propose PECC, combining complementary graph-context representations, OD-based candidate weighting, and independent OD-state graph-summary alignment for stain-calibrated local supervision.

- We propose MRSA, jointly aligning tissue-block expression distributions and intra-image OD-state relations under relaxed node-mass constraints, without fixed cross-section node assignments.

2 Related Work

2.1 Virtual Staining and Weakly Paired Translation

Virtual staining has been studied using paired and unpaired image translation (Xiong et al., 2025; Wang et al., 2024). PyramidP2P (Liu et al., 2022) mitigates local misalignment through multi-scale reconstruction while retaining coordinate-based supervision, whereas CycleGAN-based methods (Zhu et al., 2017; Zhang et al., 2022) enable unpaired translation through cycle consistency. CUT (Park et al., 2020) preserves content through contrastive learning between spatially corresponding input–output features, while ASP (Li et al., 2023) adaptively weights weakly paired positives using feature similarity. Other approaches address unreliable supervision through feature decomposition (Wang et al., 2025) or auxiliary-task alignment (Xie et al., 2026).

Transport-based methods learn soft cross-image associations: SIMGAN (Guan et al., 2025) aligns feature-level transport structures, while USIGAN (Peng et al., 2026) enforces consistency between direct and indirect transport plans under unbalanced transport. Stain-specific supervision provides complementary expression cues: PSPStain (Chen et al., 2024) incorporates focal optical density (FOD) for expression-level supervision alongside prototype-based semantic consistency.

2.2 Graph-Based Virtual Staining

Graph-based methods model contextual relations among tissue regions. GSGStain (Qiu et al., 2026) constructs an H&E-derived cell graph with morphological, spatial, and adjacent-IHC biomarker features, and rectifies noisy biomarker semantics through graph context. M2PL-GAN (Liu et al., 2026) aligns feature-similarity patch graphs through bidirectional contrastive learning, while TASTain (Jiang et al., 2026) introduces topology perturbation and node-importance-aware matching.

Our distinction is to use continuous OD compatibility jointly for feature-relation calibration and spatially proposed candidate weighting in PECC. MRSA complements these constraints by aligning tissue-block expression distributions and intra-image OD-state relations.

3 Methodology

Given an H&E input patch X and its weakly paired real-IHC reference Y, the generator produces $\hat{Y} = G(X)$ (Figure 2(a)). We use DAB optical density (OD) as a continuous staining-intensity proxy rather than a direct measurement of protein concentration. Following PSPStain (Chen et al., 2024), we obtain the focal OD field O^q, $q \in \{\hat{Y}, Y\}$, by DAB color deconvolution (Ruifrok et al., 2001) and a focal intensity transform with $\alpha = 1.8$. The exact transform is given in Appendix A.1.2. This field provides common staining evidence for calibrating local supervision in PECC and aligning regional expression states in MRSA.

3.1 Stain-Calibrated Correspondence Learning

PECC treats spatial co-location as a correspondence proposal rather than a verified match. It combines stain-conditioned contextual representations with OD-calibrated candidate supervision, complemented by alignment of OD-state graph summaries (Figure 2(b)).

At each selected encoder layer, we extract patch features $\{\mathbf{h}_i^q\}_{i=1}^N$ and corresponding FOD samples $\{o_i^q\}_{i=1}^N$. Feature affinity captures visual context but does not necessarily imply compatible staining, whereas OD similarity alone can connect visually unrelated regions. We therefore condition feature relations on OD compatibility while retaining the original feature-context graph. Using $\kappa_{\mathrm{OD}}(a, b) = \exp(-|a - b|/\sigma_{\mathrm{OD}})$ and cosine similarity $s(\cdot, \cdot)$, we define

$$A_{ij}^{F,q} = \mathbb{I}\big[s(\mathbf{h}_i^q, \mathbf{h}_j^q) > \tau\big], \qquad A_{ij}^{\mathrm{OD},q} = \mathbb{I}\big[s(\mathbf{h}_i^q, \mathbf{h}_j^q)\kappa_{\mathrm{OD}}(o_i^q, o_j^q) > \tau\big], \tag{1}$$

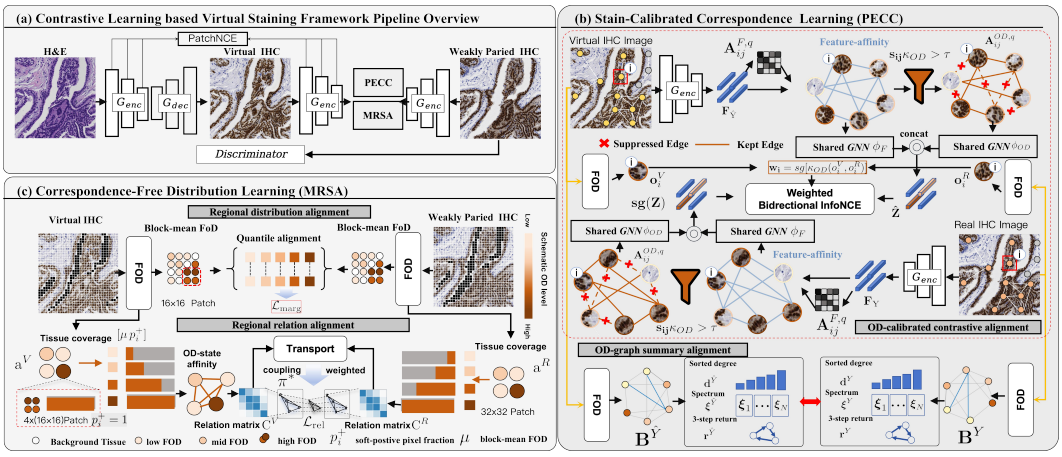

Figure 2: Overview of PGLStain. (a) Training framework. (b) PECC combines dual-graph context encoding with OD-weighted candidate supervision; independent OD-only graphs provide complementary summary alignment. (c) MRSA compares regional-mean distributions at common quantiles and within-image OD-state relations under a soft UGW coupling. Relation matrices and tissue-coverage masses determine the coupling, whose detached values weight relational comparison. V/R denote virtual/reference IHC. Nodes and quantile samples are schematic.

Since $\kappa_{\mathrm{OD}} \leq 1$ and $\tau > 0$, the OD-calibrated graph retains a subset of feature-affinity edges. The two graphs retain feature-affinity and stain-conditioned contextual views over the same nodes. After adding self-loops and applying symmetric degree normalization, we encode and concatenate these views:

$$\mathbf{z}_i^q = \Phi_F\big(\mathbf{H}^q, \widetilde{\mathbf{A}}^{F,q}\big)_i \,\|\, \Phi_{\mathrm{OD}}\big(\mathbf{H}^q, \widetilde{\mathbf{A}}^{\mathrm{OD},q}\big)_i. \tag{2}$$

Here, $\mathbf{H}^q$ stacks the sampled features. Graphs are constructed independently within each image, with each encoder shared across virtual and real IHC. The real-IHC representation is detached during generator optimization.

Contextual encoding does not by itself resolve the uncertainty of cross-section correspondence. We therefore calibrate the contribution of each spatially proposed candidate using its cross-image FOD discrepancy (Figure 2(b)). At layer ℓ, the detached weight is $w_i^\ell = \mathrm{sg}[\kappa_{\mathrm{OD}}(\hat{o}_i^\ell, o_i^\ell)]$, normalized as $\bar{w}_i^\ell = w_i^\ell / \sum_j w_j^\ell$. These weights are evaluated between images, rather than inherited from the intra-image graph edges. With sg denoting stop-gradient, the bidirectional contrastive objective is

$$\mathcal{L}_{\mathrm{corr}} = \frac{1}{|\mathcal{S}_G|} \sum_{\ell \in \mathcal{S}_G} \sum_i \bar{w}_i^\ell \Big[\ell_{\mathrm{NCE}}(\hat{\mathbf{z}}_i^\ell, \mathbf{z}_i^\ell) + \ell_{\mathrm{NCE}}(\mathbf{z}_i^\ell, \hat{\mathbf{z}}_i^\ell)\Big], \tag{3}$$

where $\mathcal{S}_G$ denotes the selected encoder layers and ℓ_{NCE} contrasts each candidate positive with other sampled nodes in the opposite image. The shared OD kernel therefore conditions contextual aggregation and allocates the relative contributions of cross-image supervision.

To complement these candidate-based comparisons, we align OD-state relations through graph-level summaries that do not require node-wise correspondence. A separate soft OD-only graph is defined at the sampled locations by $B_{ij}^q = \kappa_{\mathrm{OD}}(o_i^q, o_j^q)$. Unlike the context graphs, $\mathbf{B}^q$ is used for statistical alignment rather than feature propagation. We summarize it by its sorted normalized degrees $\mathbf{d}^q$, normalized-Laplacian spectrum $\boldsymbol{\xi}^q$, and mean three-step return statistic r^q, whose definitions are given in Appendix A.1.2. The summary alignment illustrated in Figure 2(b) minimizes

$$\mathcal{L}_{\mathrm{struct}} = \left\|\mathbf{d}^{\hat{Y}} - \mathbf{d}^{Y}\right\|_{1,\mathrm{m}} + \beta_{\mathrm{spec}} \left\|\boldsymbol{\xi}^{\hat{Y}} - \boldsymbol{\xi}^{Y}\right\|_{1,\mathrm{m}} + \beta_{\mathrm{ret}} \left|r^{\hat{Y}} - r^{Y}\right|, \tag{4}$$

where $\|\mathbf{v}\|_{1,\mathrm{m}}$ denotes the mean absolute value of its elements. These summaries are invariant to reindexing a fixed node set. The resulting PECC objective is $\mathcal{L}_{\mathrm{PECC}} = \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{struct}}\mathcal{L}_{\mathrm{struct}}$.

3.2 Correspondence-Free Distribution Learning

PECC constrains sampled local representations and their associated OD-state summaries. MRSA complements these constraints by aligning expression statistics over independently constructed tissue-block sets (Figure 2(c)). It compares regional expression composition and intra-image state relations without requiring fixed cross-section block assignments.

For marginal alignment, we independently partition each image into non-overlapping fine blocks and select them by tissue coverage using a detached intensity-threshold mask. Their tissue-masked mean FOD values form an unordered set $\{m_i^q\}_{i=1}^{N_q}$ with equal-weight empirical measure $\mu^q = N_q^{-1}\sum_i \delta_{m_i^q}$. Comparing the corresponding quantile functions yields

$$\mathcal{L}_{\mathrm{marg}} = \frac{1}{K}\sum_{k=1}^{K}\left|Q_{\hat{Y}}(u_k) - Q_Y(u_k)\right|, \qquad u_k = \frac{k-\frac{1}{2}}{K}, \tag{5}$$

where Q_q is the quantile function of μ^q. This approximates the one-dimensional Wasserstein-1 distance between distributions of regional mean expression. As illustrated by the quantile alignment in Figure 2(c), the common quantile levels specify which distribution values are compared, without pairing their spatial locations. The comparison retains regional averaging rather than reducing each image to a histogram of individual pixel intensities.

To additionally characterize relations among expression states, we independently extract coarser blocks and retain a bounded node set based on tissue coverage. Each node is described by $\mathbf{x}_i^q = [m_i^q, p_i^q]^\top$, where m_i^q and p_i^q are the tissue-masked mean FOD and soft DAB-positive pixel fraction within the coarse block. Its prescribed mass is $a_i^q = t_i^q / \sum_j t_j^q$, with tissue fraction t_i^q normalized over the retained nodes. We define the intra-image relation matrix

$$C_{ij}^q = \exp\left[-\frac{1}{2}\left\|(\mathbf{x}_i^q - \mathbf{x}_j^q) \oslash \boldsymbol{\sigma}\right\|_2^2\right], \qquad \boldsymbol{\sigma} = (\sigma_{\mathrm{OD}}, \sigma_+), \tag{6}$$

where the bandwidths control sensitivity in the two descriptor dimensions. These affinities encode similarity between regional expression states, not spatial adjacency or cell identity.

Because the two images yield independently constructed state sets, we compare their internal relations through a soft coupling $\boldsymbol{\pi}$. An entry π_{ik} associates virtual node i with reference node k, giving the relational discrepancy

$$\mathcal{E}_{\mathrm{GW}}(\boldsymbol{\pi}) = \sum_{i,j,k,l}\left|C_{ij}^{\hat{Y}} - C_{kl}^{Y}\right|^2 \pi_{ik}\pi_{jl}. \tag{7}$$

The product $\pi_{ik}\pi_{jl}$ weights the discrepancy between virtual relation (i,j) and reference relation (k,l). Thus, the coupling mediates comparison of within-image relations rather than direct cross-image descriptor distances.

To relax the prescribed node-mass constraints under weak pairing (Peng et al., 2026), we use entropic unbalanced Gromov–Wasserstein (UGW) transport and approximately estimate the coupling from

$$\min_{\boldsymbol{\pi}\geq 0} \quad \frac{1}{2}\mathcal{E}_{\mathrm{GW}}(\boldsymbol{\pi}) + \rho\,\mathrm{KL}\left(\boldsymbol{\pi}\mathbf{1}\,\|\,\mathbf{a}^{\hat{Y}}\right) + \rho\,\mathrm{KL}\left(\boldsymbol{\pi}^\top\mathbf{1}\,\|\,\mathbf{a}^{Y}\right) + \varepsilon\,\mathrm{KL}\left(\boldsymbol{\pi}\,\|\,\mathbf{a}^{\hat{Y}}\otimes\mathbf{a}^{Y}\right). \tag{8}$$

Here, KL denotes generalized KL divergence for nonnegative measures. The coefficient ρ penalizes deviations from the prescribed tissue-coverage masses, while ε controls entropic regularization. Fixed-budget GW linearizations and unbalanced Sinkhorn updates estimate $\boldsymbol{\pi}^*$; solver details are provided in the appendix.

The KL terms regularize coupling estimation rather than entering the generator's relational loss as additional penalties. For generator optimization, we detach $\boldsymbol{\pi}^*$ and minimize the normalized relational discrepancy (Figure 2(c)):

$$\mathcal{L}_{\mathrm{rel}} = \frac{\mathcal{E}_{\mathrm{GW}}(\boldsymbol{\pi}^*)}{\|\boldsymbol{\pi}^*\|_1^2}. \tag{9}$$

Squared-mass normalization removes uniform coupling-scale effects. Gradients propagate through $\mathbf{C}^{\hat{Y}}$, while the coupling and $\mathbf{C}^{Y}$ remain detached. The objective is invariant to consistent reindexing of nodes, masses, and coupling. The MRSA objective is $\mathcal{L}_{\mathrm{MRSA}} = \lambda_{\mathrm{marg}}\mathcal{L}_{\mathrm{marg}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}$.

The total generator loss is

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\mathrm{PECC}} + \mathcal{L}_{\mathrm{MRSA}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{base}}$ contains the adversarial, PatchNCE, and Gaussian-pyramid reconstruction objectives (Li et al., 2023). All proposed auxiliary objectives are used only during training.

4 Experiments

4.1 Setup

Datasets. We evaluate on MIST (Li et al., 2023) across ER, PR, HER2, and Ki67, and on the ER subset of ACROBAT 2023 (Weitz et al., 2023). All experiments use 512×512 inputs for training and inference. Dataset details and stain-specific splits are provided in Appendix A.1.3

Evaluation metrics. We assess perceptual fidelity, feature-distribution similarity, and staining agreement using eight metrics. $\mathrm{PHV}_{\mathrm{Avg}}$ (Liu et al., 2021) and DISTS (Ding et al., 2020) quantify paired perceptual differences. FID (Heusel et al., 2017) and PathFID–StainNet measure Fréchet distances in Inception and IHC- and special-stain-pretrained DINO ViT feature spaces, respectively (Li et al., 2026). HistCorr measures the correlation between normalized two-dimensional hue–saturation histograms, while DAB–KL averages the KL divergence between 256-bin histograms of the color-deconvolved DAB channels. We additionally report PearsonR between per-image mIoD values and the absolute dataset-level mean mIoD bias, $\left|N^{-1}\sum_{n=1}^{N}[\mathrm{mIoD}(\hat{Y}_n) - \mathrm{mIoD}(Y_n)]\right| \times 10^3$. Higher HistCorr and PearsonR and lower values of the remaining metrics are preferred. Implementation details and complete hyperparameters are provided in the appendix.

4.2 Comparison with Existing Methods

We compare with three general translation methods, CycleGAN (Zhu et al., 2017), CUT (Park et al., 2020), and UNSB (Kim et al., 2024), and eight IHC virtual-staining methods: PyramidP2P (Liu et al., 2022), ASP (Li et al., 2023), PPT (Zhang et al., 2024), SIMGAN (Guan et al., 2025), PSP-Stain (Chen et al., 2024), MDCL (Wang et al., 2024), M2PLGAN (Liu et al., 2026), and USI-GAN (Peng et al., 2026). All methods are implemented within a unified framework using the same data splits, input resolution, and evaluation protocol, while retaining method-specific optimization settings. Each reported result is obtained from a single training run with seed 42.

As shown in Table 1, PGLStain achieves the best values across all eight metrics on MIST–ER. Relative to the strongest competing values, DISTS decreases from 0.2100 to 0.1994 and PathFID–StainNet from 158.06 to 145.15, while HistCorr and PearsonR reach 0.7408 and 0.9455, respectively. These improvements span perceptual fidelity, feature-distribution similarity, and OD-based staining agreement. Figure 3 provides corresponding qualitative comparisons.

On ACROBAT–ER, PGLStain achieves the lowest $\mathrm{PHV}_{\mathrm{Avg}}$ and highest PearsonR, but does not lead the remaining metrics. The results indicate competitive performance with dataset- and metric-dependent advantages. Comparisons on the remaining MIST biomarkers are provided in the appendix.

Table 1: Quantitative comparisons on $\mathrm{MIST}_{\mathrm{ER}}$ and $\mathrm{ACROBAT}_{\mathrm{ER}}$. The best and second-best results on each dataset are highlighted in **bold** and underlined, respectively. PathFID uses StainNet features.

Dataset	Method	HistCorr ↑	$\mathrm{PHV}_{\mathrm{Avg}}$ ↓	DISTS ↓	FID ↓	PearsonR ↑	$\lvert\mathrm{Bias}_{\mathrm{mIoD}}\rvert$ ($\times 10^3$) ↓	DAB–KL ↓	PathFID–StainNet ↓
$\mathrm{MIST}_{\mathrm{ER}}$	CUT	0.5408	0.6074	0.2591	33.83	0.5366	3.780	1.2325	235.52
	CycleGAN	0.6303	0.5990	0.2599	44.68	0.5088	36.796	1.0273	339.50
	UNSB	0.5645	0.6296	0.2690	32.06	0.5253	5.792	1.1511	256.84
	PyramidP2P	0.5680	0.5837	0.2770	68.64	0.8936	20.216	0.2851	455.39
	ASP	0.6620	0.5562	0.2312	33.37	0.8157	9.925	0.5568	202.11
	PPT	0.6828	0.5549	0.2436	33.08	0.7217	9.071	0.7324	239.32
	PSPStain	0.6998	0.5339	0.2139	31.25	0.9203	5.288	0.4369	161.31
	MDCL	0.7162	0.5260	0.2100	30.76	0.9091	10.368	0.3874	168.16
	SIMGAN	0.6040	0.5956	0.2514	31.96	0.8098	7.308	0.8597	189.98
	USIGAN	0.6832	0.5438	0.2189	30.60	0.8675	3.339	0.6199	158.06
	M2PLGAN	0.7173	0.5576	0.2176	30.83	0.9180	6.662	0.5194	160.83
	PGLStain (Ours)	**0.7408**	**0.5057**	**0.1994**	**29.35**	**0.9455**	**1.176**	**0.2391**	**145.15**
$\mathrm{ACROBAT}_{\mathrm{ER}}$	CUT	0.6197	0.5463	0.2402	10.71	0.3381	7.854	0.5067	152.23
	CycleGAN	0.2661	0.5448	0.2296	10.50	0.3608	10.551	0.9557	191.84
	UNSB	0.3627	0.5670	0.2441	10.89	0.3389	3.640	0.8563	126.42
	PyramidP2P	**0.6583**	0.5297	0.2339	14.54	0.5110	10.680	**0.2781**	156.86
	ASP	0.4159	0.5593	0.2509	25.90	0.4370	19.616	0.5480	463.44
	PPT	0.5210	0.5183	0.2282	11.50	0.4793	14.323	0.6221	140.93
	PSPStain	0.6104	0.5076	0.2129	**7.65**	0.5301	3.774	0.3348	101.49
	MDCL	0.5240	0.5132	0.2104	9.24	0.4954	6.568	0.3956	166.60
	SIMGAN	0.6466	0.5068	0.2078	8.96	0.5120	2.301	0.3667	**97.07**
	USIGAN	0.6342	0.5053	0.2077	10.89	0.5218	**0.344**	0.3913	107.89
	M2PLGAN	0.6365	0.5011	**0.2033**	8.60	0.5255	5.059	0.3366	101.92
	PGLStain (Ours)	0.6222	**0.4984**	0.2046	8.03	**0.5369**	6.326	0.3163	110.20

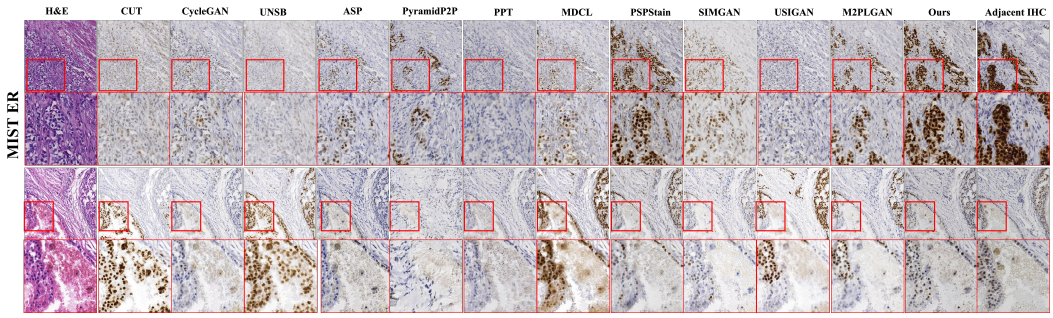

Figure 3: Qualitative comparison on MIST–ER using the same input patches and display scale across methods. Adjacent real IHC provides a weakly paired reference rather than a pixel-aligned target.

4.2.1 Pathology-Related Quantification

We evaluate Fiji-derived H-scores, positive nuclei ratios, and positive area ratios on the 1,000-patch MIST–ER validation set. Nuclear and pixel counts are pooled within each of 55 filename-defined case groups before calculating the measurements. For each method, MAE and absolute-agreement ICC(A,1) are then computed against real IHC across the same 55 paired case-level measurements, without assuming cell-wise correspondence.

Table 2 shows that PGLStain achieves the lowest MAE and highest ICC for all three measurements. Relative to the best competing MAE for each measurement, H-score error decreases from 24.18 to 18.32 (24.2%), positive nuclei ratio error from 9.84 to 8.23 percentage points (16.4%), and positive area ratio error from 5.30 to 3.84 percentage points (27.5%). These results support closer case-level agreement with weakly paired real IHC in automated staining measurements, complementing the perceptual and distributional evaluation. The measurement and aggregation protocol is detailed in Appendix A.1.4.

Table 2: Fiji-based staining agreement on 55 MIST–ER case groups. ICC denotes ICC(A,1); ratio MAEs are in percentage points.

	H-score		Positive nuclei ratio		Positive area ratio	
Method	MAE ↓	ICC ↑	MAE ↓	ICC ↑	MAE ↓	ICC ↑
CUT	84.47	0.273	31.94	0.238	16.70	0.342
CycleGAN	93.73	0.102	37.42	0.092	17.84	0.174
ASP	52.85	0.662	20.37	0.611	10.97	0.694
PPT	59.32	0.558	22.47	0.579	11.95	0.567
MDCL	32.01	0.867	14.09	0.815	7.75	0.830
PSPStain	26.52	0.925	9.84	0.913	5.30	0.914
USIGAN	29.26	0.899	13.03	0.854	6.25	0.901
M2PLGAN	24.18	0.932	10.77	0.897	5.91	0.898
PGLStain	**18.32**	**0.961**	**8.23**	**0.932**	**3.84**	**0.959**

Table 3: Objective ablation on MIST–ER. Baseline uses $\mathcal{L}_{\text{base}}$; Full combines C+S+M+R. Best values are in **bold**.

Configuration	HistCorr ↑	FID ↓	DISTS ↓	PearsonR ↑	PathFID ↓
Baseline	0.6943	32.64	0.2257	0.9095	178.65
C	0.6941	31.43	0.2261	0.9098	189.40
S	0.7027	31.04	0.2227	0.9238	170.22
PECC (C+S)	0.7031	30.66	0.2173	0.9230	169.75
MRSA (M+R)	0.7233	31.25	0.2169	0.9306	177.82
C+S+M	0.7235	31.29	0.2185	0.9368	170.83
C+S+R	0.7215	30.44	0.2164	0.9368	157.07
Full	**0.7408**	**29.35**	**0.1994**	**0.9455**	**145.15**

4.3 Ablation Study and Analysis

All ablations are conducted on MIST–ER. We denote $\mathcal{L}_{\text{corr}}$, $\mathcal{L}_{\text{struct}}$, $\mathcal{L}_{\text{marg}}$, and $\mathcal{L}_{\text{rel}}$ by C, S, M, and R, respectively. PECC comprises C+S, whereas MRSA comprises M+R. Configurations containing R in Table 3 use $\rho = 2$. PathFID denotes PathFID–StainNet throughout.

4.3.1 Effectiveness of PECC and MRSA

Table 3 examines the individual and combined contributions of the four objectives. C alone lowers FID but leaves HistCorr, DISTS, and PearsonR nearly unchanged and increases PathFID. In contrast, S, which aligns OD-state graph summaries, improves HistCorr, PearsonR, and PathFID. Combining C and S improves all five metrics over the baseline, supporting their joint use in PECC. MRSA (M+R) also improves all five metrics, with larger gains in HistCorr and PearsonR than PECC alone. It therefore provides a complementary means of learning expression-state distributions and relations without fixed cross-section node assignments.

Adding R to PECC lowers PathFID from 169.75 to 157.07, whereas adding M does not improve this metric. This contrast is consistent with distinct contributions from marginal composition alignment and relational matching. Full achieves the best values across all five reported metrics among the listed configurations, supporting the joint formulation under the evaluated protocol. Qualitative module ablations are provided in Appendix A.1.6.

4.3.2 Stain-Calibrated Graph Context and Weighting

Feature similarity commonly guides graph construction and candidate weighting (Li et al., 2023; Liu et al., 2026; Jiang et al., 2026), but shared morphological context does not necessarily imply compatible staining states. We therefore examine graph context and candidate weighting separately, holding the other component fixed (Table 4).

For the topology comparison, all variants retain continuous OD weighting and two independently parameterized encoder branches with matched capacity. Node-wise uses identity adjacencies, removing cross-node aggregation while retaining learnable transformations. Feature-only and OD-calibrated-only supply the corresponding adjacency to both branches; Full combines the two views. The positive candidates, remaining objectives, and training settings are unchanged.

Full improves all five metrics over Node-wise, whereas Feature-only performs worse on every metric. OD-calibrated-only improves all metrics over Feature-only. Combining both views further improves DISTS, FID, DAB–KL, and PathFID, with slightly lower HistCorr than OD-calibrated-only. These comparisons support complementary feature and stain-calibrated context under OD-aware weighting, rather than an unconditional benefit from graph propagation.

We next fix the dual topology and vary only candidate weighting: uniform, detached feature similarity, hard OD gating, or continuous OD weighting. All variants retain the same spatially proposed

Table 4: Graph-context and candidate-weighting ablations on MIST–ER. Topology controls share continuous OD weighting (Cont.); weighting controls share the dual topology. OD-calib. denotes the OD-calibrated-only control. Full is their shared reference. Best values across all listed variants are in **bold**.

Topology	Weight	HistCorr↑	DISTS↓	FID↓	DAB–KL↓	PathFID↓
Node-wise	Cont.	0.7316	0.2100	30.23	0.2736	164.01
Feature-only	Cont.	0.7187	0.2137	30.99	0.3318	170.07
OD-calib.	Cont.	**0.7422**	0.2081	30.44	0.2605	152.57
Dual	Uniform	0.7221	0.2085	30.49	0.2813	165.53
Dual	Feature similarity	0.7320	0.2131	30.75	0.3064	166.80
Dual	Hard OD	0.7140	0.2118	30.46	0.3246	153.07
Dual (Full)	**Cont.**	0.7408	**0.1994**	**29.35**	**0.2391**	**145.15**

Table 5: Alternatives to $\mathcal{L}_{\mathrm{rel}}$ on MIST–ER, with identical OD-state descriptors, node construction, and remaining training settings. UGW uses $\rho = 2.00$. Best and second-best values are **bold** and underlined.

Formulation	HistCorr↑	FID↓	PearsonR↑	DAB–KL↓	PathFID↓
Full w/o $\mathcal{L}_{\mathrm{rel}}$	0.7235	31.29	0.9368	0.3103	170.83
Direct 2-D UOT	0.7330	30.92	0.9402	0.2747	<u>151.08</u>
Joint MMD	0.7286	30.44	0.9403	0.2983	153.82
BGW	<u>0.7357</u>	<u>30.15</u>	<u>0.9410</u>	<u>0.2539</u>	157.86
UGW	**0.7408**	**29.35**	**0.9455**	**0.2391**	**145.15**

positive pairs and remaining training settings. Relative to uniform weighting, feature-similarity weighting improves HistCorr but worsens the other four displayed metrics. Hard gating improves FID and PathFID but worsens HistCorr, DISTS, and DAB–KL. Continuous OD weighting achieves the best values within this fixed-topology comparison, favoring gradual stain-aware attenuation over the tested feature-based and binary alternatives. Complete weighting results, including PearsonR, and mechanism visualizations are provided in Appendix A.1.5.

4.3.3 Effectiveness of the Relational Matching Formulation

We examine whether direct expression-state alignment is sufficient and whether relational matching benefits from relaxed marginals. All matching variants share the OD-state descriptors $[\mathrm{mean}(\mathrm{FOD}), p^{+}]$, node construction, auxiliary objectives, data split, and training settings, differing only in the formulation replacing $\mathcal{L}_{\mathrm{rel}}$. The Full model without this term serves as a common reference.

Direct 2-D UOT matches cross-image descriptors using a bounded RBF dissimilarity, without comparing pairwise intra-image relations. Joint MMD aligns the joint descriptor distribution using maximum mean discrepancy. BGW retains relational matching under exact marginal constraints, whereas UGW permits deviations from the prescribed node masses to accommodate differences in expression-state proportions.

As shown in Table 5, Direct 2-D UOT and Joint MMD improve all five metrics over the non-relational reference, supporting the benefit of direct expression-state alignment. UGW further improves all five metrics over both alternatives, supporting pairwise OD-state relational matching beyond these direct-alignment formulations. Its consistent improvement over BGW further supports marginal relaxation under the evaluated protocol. The controls share the same descriptor bandwidths and replacement-loss coefficient; their kernel definitions, normalization, and solver settings are detailed in Appendix A.1.7.

Additional analyses. Sensitivity to the marginal-relaxation parameter ρ, entropic coefficient ε, and loss coefficients is reported in Appendices A.1.8, A.1.9, and A.1.11, respectively. Appendix A.3 further examines frozen-model loss responses to DAB-tile permutation. This diagnostic probes sensitivity to staining rearrangement, rather than establishing correspondence correctness or invariance to image perturbations.

5 Conclusion

We presented PGLStain for OD-guided, weakly paired IHC virtual staining. PECC calibrates local supervision and aligns OD-state summaries, while MRSA matches regional distributions and relations without fixed cross-section assignments. These training-only objectives add no inference overhead. Experiments show competitive perceptual quality and staining agreement.

Ethics statement

This study performs secondary analysis of publicly released histopathology datasets under their stated access conditions. No new patient samples were collected, and the authors did not access direct patient identifiers. We rely on the ethics, consent, and data-governance procedures documented by the original dataset providers. PGLStain is developed for methodological research and is not intended for direct clinical diagnosis or treatment decisions. Virtual staining models may inherit dataset-specific biases and may fail under shifts in tissue type, patient population, specimen preparation, scanner, or staining protocol. Any clinical deployment would therefore require independent multi-center validation, prospective evaluation, and appropriate regulatory and professional oversight.

Reproducibility statement

All experiments were conducted within PuzzleStain, an in-house, configuration-driven framework that provides unified data loading, preprocessing, training, inference, and evaluation. All comparisons use consistent data partitions, image resolution, preprocessing, and metric implementations; controlled ablations additionally share the same architecture, training schedule, and random seed unless otherwise stated. Complete configurations, hyperparameters, and implementation details are documented in the appendix, and an anonymized snapshot of the experimental codebase is included in the supplementary material.

AI Use Disclosure

Generative AI tools were used to draft initial versions of manuscript sections, provide reviewer-style feedback that helped identify gaps in the experimental evaluation and informed additional experiments, and refine the language and LaTeX presentation of the manuscript. The authors take full responsibility for the final manuscript, experimental results, and conclusions.

References

Emma H. Allott, Joseph Geradts, Xuezheng Sun, Stephanie M. Cohen, Gary R. Zirpoli, Thaer Khoury, Wiam Bshara, Mengjie Chen, Mark E. Sherman, Julie R. Palmer, Christine B. Ambrosone, Andrew F. Olshan, and Melissa A. Troester. Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification. *Breast Cancer Research*, 18:68, 2016. doi: 10.1186/s13058-016-0725-1.

Fuqiang Chen, Ranran Zhang, Boyun Zheng, Yiwen Sun, Jiahui He, and Wenjian Qin. Pathological Semantics-Preserving Learning for HE-to-IHC Virtual Staining . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15004. Springer Nature Switzerland, October 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Xianchao Guan, Zheng Zhang, Yifeng Wang, Yueheng Li, and Yongbing Zhang. Supervised information mining from weakly paired images for breast ihc virtual staining. *IEEE Transactions on Medical Imaging*, 44(5):2120–2130, 2025. doi: 10.1109/TMI.2024.3525299.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Mingzhou Jiang, Jiaying Zhou, Nan Zeng, Mickael Li, Qijie Tang, Chao He, Huazhu Fu, and Honghui He. Topology-aware pathological consistency matching for weakly-paired ihc virtual staining, 2026. URL https://arxiv.org/abs/2601.02806.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *International Conference on Learning Representations*, volume 2024, pp. 19312–19331, 2024.

Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 632–641, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43987-2.

Jiawen Li, Jiali Hu, Xitong Ling, Yongqiang Lv, Yuxuan Chen, Yizhi Wang, Tian Guan, Yifei Liu, and Yonghong He. Stainnet: Scaling self-supervised foundation models on immunohistochemistry and special stains for computational pathology, 2026. URL https://arxiv.org/abs/2512.10326.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1815–1824, June 2022.

Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE transactions on medical imaging*, 40(8):1977–1989, 2021.

Zequn Liu, Liangkuan Zhu, Yining Xie, Xiaoqing Hu, Ziyu Zhang, Jing Zhao, Haochen Qi, Jiajun Chen, and Jiayi Ma. M2pl-gan: Multi-view multi-level pathology semantic perception learning for he-to-ihc virtual staining. *IEEE Transactions on Medical Imaging*, 45(6):3082–3097, 2026. doi: 10.1109/TMI.2026.3668248.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature machine intelligence*, 6(9):1077–1093, 2024.

Yue Peng, Bing Xiong, Fuqiang Chen, Deboch Eyob Abera, RanRan Zhang, Wanming Hu, Jing Cai, and Wenjian Qin. Usigan: Unbalanced self-information feature transport for weakly paired image ihc virtual staining. *IEEE Transactions on Image Processing*, 35:4729–4743, 2026. doi: 10.1109/TIP.2026.3679993.

Fanhao Qiu, Yangyang Zhang, and Zhengxia Wang. Graph-semantic guided learning for virtual immunohistochemistry staining on consecutive histology sections. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8556–8564, 2026.

Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

Song Wang, Zhong Zhang, Huan Yan, Ming Xu, and Guanghui Wang. Mix-domain contrastive learning for unpaired he-to-ihc stain translation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 2982–2988, 2024. doi: 10.1109/ICIP51287.2024.10648270.

Tong Wang, Mingkang Wang, Zhongze Wang, Hongkai Wang, Qi Xu, Fengyu Cong, and Hongming Xu. Oda-gan: Orthogonal decoupling alignment gan assisted by weakly-supervised learning for virtual immunohistochemistry staining. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25920–25929, 2025. doi: 10.1109/CVPR52734.2025.02414.

Philippe Weitz, Masi Valkonen, Leslie Solorzano, Circe Carr, Kimmo Kartasalo, Constance Boissin, Sonja Koivukoski, Aino Kuusela, Dusan Rasic, Yanbo Feng, et al. A multi-stain breast cancer histological whole-slide-image data set from routine diagnostics. *Scientific data*, 10(1):562, 2023.

Shigeng Xie, Hongming Xu, Guiyang Jiang, Tuomo Rossi, Tommi Kärkkäinen, and Fengyu Cong. Virtual immunohistochemistry staining with dual-aligned multi-task feature guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35311–35320, 2026.

Bing Xiong, Yue Peng, RanRan Zhang, Fuqiang Chen, JiaYe He, and Wenjian Qin. Unpaired multi-domain histopathology virtual staining using dual path prompted inversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8780–8787, 2025.

Ranran Zhang, Yankun Cao, Yujun Li, Zhi Liu, Jianye Wang, Jiahuan He, Chenyang Zhang, Xiaoyu Sui, Pengfei Zhang, Lizhen Cui, et al. Mvfstain: multiple virtual functional stain histopathology images generation based on specific domain mapping. *Medical Image Analysis*, 80:102520, 2022.

Wei Zhang, Tik Ho Hui, Pui Ying Tse, Fraser Hill, Condon Lau, and Xinyue Li. High-resolution Medical Image Translation via Patch Alignment-Based Bidirectional Contrastive Learning . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15004. Springer Nature Switzerland, October 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

A Appendix

A.1 Supplementary Experimental Details

A.1.1 Implementation and Reproducibility Details

PGLStain is implemented in PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. It uses a six-block ResNet generator and a five-layer PatchGAN discriminator, optimized with Adam ($\beta_1 = 0.5, \beta_2 = 0.999$). Identity PatchNCE is enabled ($\mathrm{nce}_{\mathrm{idt}} = \mathrm{true}$), and synchronized horizontal flipping is applied during training.

For MIST, all methods use a 512×512 input resolution and a 100-epoch training schedule, while retaining method-specific optimization settings from their original implementations. The initial generator learning rate is 2×10^{-4}, except for PSPStain, which uses 1×10^{-4} with a discriminator learning rate of 5×10^{-6}. Learning rates are held constant for the first 50 epochs and linearly decayed to zero over the remaining 50 epochs. Batch size is one, except for SIMGAN and USIGAN, which use two. Unless otherwise specified, all experiments use random seed 42.

Checkpoint selection. We evaluate the checkpoint saved at the final scheduled training epoch: epoch 100 for MIST and epoch 20 for ACROBAT. Intermediate checkpoints are not selected according to validation performance. This final-epoch rule is applied consistently to the method comparisons and controlled ablations.

PECC settings. The base objective uses $\lambda_{\mathrm{GAN}} = \lambda_{\mathrm{NCE}} = 1$ and $\lambda_{\mathrm{GP}} = 10$. PatchNCE samples 256 features from encoder layers $\{0, 4, 8, 12, 16\}$ with temperature 0.07. PECC operates on layers $\{8, 12, 16\}$ using four-hop graph propagation, a feature-similarity threshold of 0.6, 256 sampled nodes, and $\sigma_{\mathrm{OD}} = 0.1$. The coefficients of $\mathcal{L}_{\mathrm{corr}}$ and $\mathcal{L}_{\mathrm{struct}}$ are both 0.1. Within $\mathcal{L}_{\mathrm{struct}}$, the degree, spectrum, and three-step return summaries are weighted by 1.0, 0.5, and 0.01, respectively.

MRSA settings. Marginal alignment uses 16×16-pixel blocks, 128 quantiles, a tissue-mask intensity threshold of 0.95, a minimum tissue fraction of 0.25, and a minimum retained set size of eight blocks. Relational alignment uses 32×32-pixel blocks and retains between 8 and 64 nodes. The positive-OD threshold and temperature are 0.02 and 0.01, respectively. We use $\sigma_{\mathrm{OD}} = 0.1$, $\sigma_+ = 0.25$, and $\varepsilon = 0.05$. The marginal relaxation coefficient is $\rho = 2$. The coefficients of $\mathcal{L}_{\mathrm{marg}}$ and $\mathcal{L}_{\mathrm{rel}}$ are 5.0 and 1.0, respectively. The UGW solver performs six GW linearizations, each with 30 Sinkhorn iterations.

Block construction and tissue selection. Local blocks are non-overlapping and extracted independently from the generated and reference images. For RGB values $I^q(u) \in [0, 1]^3$, the detached tissue mask is

$$M^q(u) = \mathbb{I}\left[\frac{1}{3}\sum_{c=1}^{3} I_c^q(u) < 0.95\right]. \tag{11}$$

Algorithm 1 OD-State Unbalanced Gromov–Wasserstein Alignment

Require: Virtual IHC $\hat{Y}$; weakly paired real IHC Y
Require: Scales $\boldsymbol{\sigma} = (\sigma_{\mathrm{OD}}, \sigma_{+})$; relaxation ρ; entropy ε
Require: GW iterations T_{GW}; Sinkhorn iterations T_{SK}
Ensure: Relational loss $\mathcal{L}_{\mathrm{rel}}$

Independently extract tissue-block nodes $\{(\mathbf{x}_i, a_i)\}_{i=1}^{n}$ from $\hat{Y}$ and $\{(\mathbf{y}_j, b_j)\}_{j=1}^{m}$ from Y using detached tissue masks.
Each node contains the tissue-masked mean FOD and soft-positive pixel fraction of one retained block.
Normalize tissue-coverage masses: $\sum_i a_i = \sum_j b_j = 1$.
$C^{\hat{Y}}_{ik} \leftarrow \exp\Big[-\frac{1}{2}\,\|(\mathbf{x}_i - \mathbf{x}_k) \oslash \boldsymbol{\sigma}\|_2^2\Big]$.
$C^{Y}_{jl} \leftarrow \exp\Big[-\frac{1}{2}\,\|(\mathbf{y}_j - \mathbf{y}_l) \oslash \boldsymbol{\sigma}\|_2^2\Big]$.
Detach the reference relation matrix C^{Y}.
Estimate the coupling without gradient tracking:
Initialize $\boldsymbol{\pi} \leftarrow \mathbf{a}\mathbf{b}^{\top}$ and $\eta \leftarrow \rho/(\rho + \varepsilon)$.
for $t = 1$ to T_{GW} **do**
  $\mathbf{r} \leftarrow \boldsymbol{\pi}\mathbf{1}$, $\mathbf{c} \leftarrow \boldsymbol{\pi}^{\top}\mathbf{1}$.
  $K_1 \leftarrow (C^{\hat{Y}} \odot C^{\hat{Y}})\mathbf{r}\mathbf{1}^{\top}$.
  $K_2 \leftarrow \mathbf{1}\big[(C^{Y} \odot C^{Y})\mathbf{c}\big]^{\top}$.
  $K_3 \leftarrow 2C^{\hat{Y}}\boldsymbol{\pi}(C^{Y})^{\top}$.
  $K \leftarrow K_1 + K_2 - K_3$.
  $G \leftarrow (\mathbf{a}\mathbf{b}^{\top}) \odot \exp(-K/\varepsilon)$.
  Initialize $\mathbf{u} \leftarrow \mathbf{1}$, $\mathbf{v} \leftarrow \mathbf{1}$.
  for $s = 1$ to T_{SK} **do**
    $\mathbf{u} \leftarrow \big[\mathbf{a} \oslash (G\mathbf{v})\big]^{\eta}$.
    $\mathbf{v} \leftarrow \big[\mathbf{b} \oslash (G^{\top}\mathbf{u})\big]^{\eta}$.
  end for
  $\boldsymbol{\pi} \leftarrow \mathrm{diag}(\mathbf{u})G\,\mathrm{diag}(\mathbf{v})$.
end for
Detach the estimated coupling $\boldsymbol{\pi}$.
Recompute K as above, with gradient tracking through $C^{\hat{Y}}$ only.
return $\mathcal{L}_{\mathrm{rel}} = \langle K, \boldsymbol{\pi}\rangle / \|\boldsymbol{\pi}\|_1^2$.

This brightness-based mask excludes near-white background and does not provide cell or tissue-compartment segmentation. For block B_i, tissue coverage is $t_i^q = |B_i|^{-1}\sum_{u\in B_i} M^q(u)$. Blocks with $t_i^q \geq 0.25$ are retained. If fewer than eight qualify, the eight blocks with the highest coverage are used instead, even if some fall below the threshold. Marginal alignment retains all selected blocks; relational alignment retains only the 64 highest-coverage blocks when this limit is exceeded. The two images may therefore yield different set sizes. Block coordinates are not used for cross-image matching.

Expression statistics and gradient flow. The block mean FOD and soft-positive fraction are computed over tissue pixels. The latter averages the pixel-wise score $\mathrm{sigmoid}((O^q(u) - 0.02)/0.01)$ and is not a positive-cell fraction. Both statistics are implemented by pooling their tissue-masked numerators and dividing by $\max(t_i^q, 10^{-6})$. Relational node masses are proportional to $\max(t_i^q, 10^{-6})$ and normalized over the retained nodes. Tissue masks, reference-image quantities, and estimated couplings are detached. Gradients pass through the generated-image FOD statistics and relation matrix, but not through node selection or transport iterations.

Transport conventions and numerical safeguards. We use generalized KL divergence,

$$\mathrm{KL}(\mathbf{u}\|\mathbf{v}) = \sum_r \left[u_r \log \frac{u_r}{v_r} - u_r + v_r\right], \tag{12}$$

and define $\|\boldsymbol{\pi}\|_1 = \sum_{i,k} \pi_{ik}$. Divisions, powers, and exponentials in Algorithm 1 are element-wise unless matrix multiplication is indicated. The linearized GW cost is clamped to be nonnegative. Exponential factors and coupling entries are floored at 10^{-30}; Sinkhorn denominators are floored at 10^{-12}, and scaling ratios are clipped to $[10^{-8}, 10^{8}]$ before exponentiation. The transported mass is floored at 10^{-8} before squaring in the loss denominator. The pseudocode omits these numerical safeguards for readability. The fixed iteration budget provides an approximate coupling, not a guarantee of convergence.

A.1.2 FOD AND OD-STATE SUPERVISION DETAILS

Focal optical density. Following the Beer–Lambert relation $\mathrm{OD} = -\log_{10}(I/I_0)$, color deconvolution separates the DAB component and reconstructs its grayscale transmittance g^q_{DAB} (Chen et al., 2024; Ruifrok et al., 2001). We use the focal transform

$$O^q(\mathbf{u}) = \left[\log_{10} \frac{1}{g^q_{\mathrm{DAB}}(\mathbf{u}) + \delta_\alpha}\right]_+^\alpha, \quad \delta_\alpha = 10^{-\exp(1/\alpha)}, \tag{13}$$

where $q \in \{\hat{Y}, Y\}$, $[x]_+ = \max(x, 0)$, and $\alpha = 1.8$ follows PSPStain (Chen et al., 2024). The FOD field is a transformed staining-intensity proxy; it is not a direct estimate of protein concentration.

Shared OD kernel and candidate weights. PECC uses the same scalar kernel for intra-image graph calibration and cross-image candidate weighting:

$$\kappa_{\mathrm{OD}}(a, b) = \exp\left(-\frac{|a - b|}{\sigma_{\mathrm{OD}}}\right), \tag{14}$$

with bandwidth σ_{OD}. For spatially proposed candidates at encoder layer ℓ, the detached weights and their normalization are

$$w_i^\ell = \mathrm{sg}\left[\kappa_{\mathrm{OD}}\left(\hat{o}_i^\ell, o_i^\ell\right)\right], \qquad \bar{w}_i^\ell = \frac{w_i^\ell}{\sum_j w_j^\ell}, \tag{15}$$

where sg stops gradients. The graph construction applies this kernel between nodes of one image, whereas the candidate weights compare virtual and reference FOD values. They are separate computations, not edge weights transferred between graphs.

Independent OD-graph summaries. The summary branch uses the full soft OD-only graph

$$B_{ij}^q = \kappa_{\mathrm{OD}}(o_i^q, o_j^q), \qquad q \in \{\hat{Y}, Y\}. \tag{16}$$

The nodes use the same sampled locations as the context branches. This graph is independent of the feature-propagation graphs. Its degree matrix, normalized Laplacian, and transition matrix are

$$\mathbf{D}^q = \mathrm{diag}(\mathbf{B}^q\mathbf{1}), \qquad \mathbf{L}^q = \mathbf{I} - (\mathbf{D}^q)^{-1/2}\mathbf{B}^q(\mathbf{D}^q)^{-1/2}, \qquad \mathbf{P}^q = (\mathbf{D}^q)^{-1}\mathbf{B}^q, \tag{17}$$

and the corresponding summaries are

$$\mathbf{d}^q = \mathrm{sort}\left(\frac{\mathbf{B}^q\mathbf{1}}{N}\right), \qquad \boldsymbol{\xi}^q = \mathrm{eigvalsh}(\mathbf{L}^q), \qquad r^q = \frac{1}{N}\,\mathrm{tr}\left((\mathbf{P}^q)^3\right). \tag{18}$$

Here, eigvalsh returns the ordered eigenvalues of the symmetric normalized Laplacian. The degree summary records the distribution of normalized affinity strengths, while the spectrum and return statistic summarize the OD-state graph's spectral structure. In particular, r^q is a nonlinear spectral summary, not an independent spatial topology descriptor. Sorting, eigenvalues, and the trace are invariant to consistent reindexing of a fixed node set; this does not imply invariance to image-content perturbations that change the node values.

Table 6: MIST splits used in our experiments. Counts denote matched H&E–IHC patch pairs.

Target stain	Training pairs	Evaluation pairs
ER	4,153	1,000
PR	4,139	1,000
HER2	4,642	1,000
Ki67	4,361	1,000

Regional descriptors and prescribed masses. Relational alignment uses independently selected coarse tissue blocks:

$$\mathbf{x}_i^q = \begin{bmatrix} m_i^q & p_i^q \end{bmatrix}^\top, \qquad a_i^q = \frac{t_i^q}{\sum_j t_j^q}, \tag{19}$$

Here, m_i^q and p_i^q are computed from the pixels within the retained coarse block under its tissue mask. The latter is a soft-positive pixel fraction, not a fraction of positive fine blocks or detected nuclei. The mass a_i^q normalizes tissue fraction t_i^q over the retained node set. By contrast, marginal alignment uses an equal-weight empirical distribution of retained fine-block means.

Module objectives. For completeness, the two module objectives are

$$\mathcal{L}_{\mathrm{PECC}} = \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{struct}}\mathcal{L}_{\mathrm{struct}}. \tag{20}$$

$$\mathcal{L}_{\mathrm{MRSA}} = \lambda_{\mathrm{marg}}\mathcal{L}_{\mathrm{marg}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}. \tag{21}$$

These objectives enter the total generator objective in Eq. 10. The loss coefficients and the tissue-selection and transport settings are specified in the implementation details.

A.1.3 Datasets and Preprocessing

MIST. MIST (Li et al., 2023) provides weakly paired H&E–IHC images for ER, PR, HER2, and Ki67. We follow the publicly released benchmark training/validation partition. We follow the publicly released patch-level training/validation partition. All methods use the same partition. All methods use the same partition.

Each target stain is evaluated on its fixed validation set of 1,000 matched H&E–IHC patch pairs. For MIST–ER, these validation patches originate from 55 WSIs and represent sampled regions rather than complete coverage of those slides. All subsequent MIST–ER validation analyses use this same 1,000-pair set. Where results are aggregated by WSI, grouping follows the source-WSI identifiers and includes only the available validation patches from each slide. Such summaries therefore describe the sampled validation regions, not measurements over complete WSIs.

We evaluate all four target stains, with ER comparisons presented in the main text and additional quantitative and qualitative results for PR, HER2, and Ki67 provided in the appendix. Table 6 summarizes the stain-specific training and validation partitions. All counts denote filename-matched H&E–IHC patch pairs.

ACROBAT 2023. For ACROBAT 2023 (Weitz et al., 2023), we evaluate the ER subset using patches from 382 training WSIs and 44 evaluation WSIs. The $1{,}024 \times 1{,}024$ source patches are resized to 512×512 for both training and inference, matching the input resolution used in the MIST experiments. For ACROBAT–ER, the training and evaluation partitions contain disjoint source WSIs, comprising 382 training WSIs and 44 evaluation WSIs.

A.1.4 Fiji-Based Staining Quantification

Evaluation protocol. We quantify H-scores, positive nuclei ratios, and positive area ratios on the 1,000 filename-matched MIST–ER validation patches. All methods use the same H&E inputs and

real-IHC references. A fixed Fiji-based pipeline is applied separately to each image source, with identical preprocessing, detection settings, stain vectors, and intensity thresholds across methods. Nuclear and pixel counts are pooled within 55 filename-defined case groups before computing the measurements. Each method is compared separately with real IHC using the same groups, without assuming cell-wise correspondence between adjacent sections.

Image processing and staining classification. The 512×512 RGB patches are bilinearly resized to 1024×1024. H–DAB colour deconvolution separates the hematoxylin and DAB components, and staining intensity is measured from the 8-bit DAB channel, where lower values indicate stronger staining. Nuclear ROIs are detected independently in each image using the StarDist *Versatile (H&E nuclei)* model followed by mask post-processing. Contrast enhancement is applied only to the segmentation input; staining intensities are measured from the unenhanced deconvolution output.

Each nuclear ROI is classified by its mean DAB-channel intensity $\bar{I}$ as strong ($\bar{I} \leq 60$), moderate ($60 < \bar{I} \leq 120$), weak ($120 < \bar{I} \leq 180$), or negative ($180 < \bar{I} \leq 235$). ROIs with $\bar{I} > 235$ are excluded by the fixed background rule. These criteria are identical for real IHC and all generated outputs.

Staining measurements and aggregation. Counts are aggregated separately for each method and the real-IHC reference; the image-source index is omitted below. Let $n_{i,k}$ denote the number of retained nuclei in patch i assigned to category $k \in \{0, 1, 2, 3\}$, corresponding to negative, weak, moderate, and strong staining, respectively. For the patch set $\mathcal{P}_c$ of case group c, define

$$N_{c,k} = \sum_{i \in \mathcal{P}_c} n_{i,k}, \qquad N_c = \sum_{k=0}^{3} N_{c,k}. \tag{22}$$

For $N_c > 0$, the H-score and positive nuclei ratio are

$$H_c = 100\,\frac{N_{c,1} + 2N_{c,2} + 3N_{c,3}}{N_c}, \qquad P_c^{\mathrm{nuc}} = 100\,\frac{N_{c,1} + N_{c,2} + N_{c,3}}{N_c}. \tag{23}$$

The H-score ranges from 0 to 300, while the positive nuclei ratio is expressed as a percentage.

Positive area is measured pixelwise from the same DAB channel. For patch i, let A_i^{+} count pixels with intensity $I \leq 180$, and let A_i^{eval} count pixels with $I \leq 235$. The positive area ratio is

$$P_c^{\mathrm{area}} = 100\,\frac{\sum_{i \in \mathcal{P}_c} A_i^{+}}{\sum_{i \in \mathcal{P}_c} A_i^{\mathrm{eval}}}. \tag{24}$$

The denominator is an intensity-defined evaluated area. All three measurements are calculated from pooled counts, rather than by averaging patch-level scores or percentages.

Accuracy and agreement. For measurement $z \in \{H, P^{\mathrm{nuc}}, P^{\mathrm{area}}\}$, let $\hat{z}_c^{(m)}$ and z_c^{ref} denote the values for method m and real IHC, respectively. Mean absolute error is computed as

$$\mathrm{MAE}_{m,z} = \frac{1}{55} \sum_{c=1}^{55} \left| \hat{z}_c^{(m)} - z_c^{\mathrm{ref}} \right|. \tag{25}$$

Each case group receives equal weight. H-score MAE is reported in H-score units, and ratio MAEs are reported in percentage points.

Absolute agreement is assessed using McGraw–Wong ICC(A,1), the two-way mixed-effects, absolute-agreement, single-measure intraclass correlation coefficient. For each method and measurement, ICC is computed from a 55×2 matrix containing the paired method and real-IHC values. These metrics summarize case-group-level staining agreement under the common automated quantification protocol.

A.1.5 PECC Graph Context and Candidate Weighting

Graph propagation introduces intra-image context, but feature affinity does not necessarily imply staining compatibility. We therefore examine graph context and candidate weighting separately, holding the other component fixed. The topology controls in Table 4 retain continuous

OD weighting and two independently parameterized encoder branches with matched capacity and embedding dimensionality. Node-wise uses identity adjacencies, retaining learnable transformations without cross-node aggregation. Feature-only supplies the feature-affinity adjacency to both branches, whereas OD-calibrated-only supplies the OD-calibrated adjacency to both. Full combines one branch of each type. Positive-pair indices, remaining objectives, data partitions, and training settings are unchanged.

Full improves all five metrics over Node-wise, whereas Feature-only performs worse on every metric. OD-calibrated-only improves all metrics over Feature-only, but does not uniformly outperform Node-wise: its FID is 30.44 versus 30.23. Combining the two graph views further improves DISTS, FID, DAB–KL, and PathFID over OD-calibrated-only, with slightly lower HistCorr (0.7408 versus 0.7422). These results support complementary feature-context and stain-calibrated representations under OD-aware weighting, rather than an unconditional benefit from graph propagation.

To examine candidate weighting independently of graph context, we fix the dual topology and vary only the relative contribution of each spatially proposed positive pair. At encoder layer ℓ, let $\hat{\mathbf{f}}_i^\ell$ and $\mathbf{f}_i^\ell$ denote the corresponding ℓ_2-normalized projected features before graph aggregation, and let $\hat{o}_i^\ell$ and o_i^ℓ denote their sampled FOD values. The cross-image FOD discrepancy and feature similarity are

$$\begin{aligned} d_i^\ell &= \left|\hat{o}_i^\ell - o_i^\ell\right|, \\ c_i^\ell &= \left(\hat{\mathbf{f}}_i^\ell\right)^\top \mathbf{f}_i^\ell. \end{aligned} \tag{26}$$

We compare four weighting rules:

$$\begin{aligned} w_{i,\text{uniform}}^\ell &= 1, \\ w_{i,\text{feature}}^\ell &= \text{clip}_{[0,1]}\left(\frac{1+c_i^\ell}{2}\right), \\ w_{i,\text{hard}}^\ell &= \mathbb{I}\left[d_i^\ell \le \delta_{\text{hard}}\right], \\ w_{i,\text{continuous}}^\ell &= \exp\left(-\frac{d_i^\ell}{\sigma_{\text{OD}}}\right). \end{aligned} \tag{27}$$

Uniform weighting treats all candidates equally. Feature-similarity weighting maps cosine similarity to a nonnegative contribution, whereas the two OD-based rules use the same staining-discrepancy signal through either binary gating or gradual attenuation. Continuous weighting uses $\sigma_{\text{OD}} = 0.1$.

For each rule, weights are detached and normalized within each layer:

$$\bar{w}_i^\ell = \frac{\text{sg}\left(w_i^\ell\right)}{\max\left\{\sum_j \text{sg}\left(w_j^\ell\right), 10^{-8}\right\}}, \tag{28}$$

where sg denotes stop-gradient. The same normalized weights are applied to the per-query losses in both contrastive directions, and the resulting losses are averaged across the selected encoder layers. If all hard-gate weights are zero, the corresponding layer contributes zero correspondence loss; no uniform-weight fallback is used. Weighting changes neither the positive-pair indices nor the negative candidate sets. Graph encoders, auxiliary objectives, and the remaining training settings are shared across all variants.

The hard threshold $\delta_{\text{hard}} = 0.12024$ was calibrated before training the hard-gate variant using Full-model outputs for 128 MIST–ER validation image pairs. It was chosen so that the retained-pair fraction approximated the normalized effective sample size of the continuous weights on the calibration sample. For N_{cal} sampled candidates with continuous weights v_j, this quantity is

$$\eta_{\text{continuous}} = \frac{\left(\sum_{j=1}^{N_{\text{cal}}} v_j\right)^2}{N_{\text{cal}} \sum_{j=1}^{N_{\text{cal}}} v_j^2}. \tag{29}$$

The calibrated threshold remained fixed throughout the hard-gate run. Table 7 reports the complete results, including PearsonR omitted from the compact main-text comparison.

Table 7: Complete candidate-weighting results on MIST–ER. All variants retain the dual topology, auxiliary objectives, and training settings, differing only in the weighting rule. PearsonR is reported in addition to the five metrics in Table 4. PathFID uses StainNet features; best values are in **bold**.

Weighting	HistCorr ↑	DISTS ↓	FID ↓	PearsonR ↑	DAB–KL ↓	PathFID ↓
Uniform weighting	0.7221	0.2085	30.49	0.9393	0.2813	165.53
Feature-sim. weighting	0.7320	0.2131	30.75	0.9393	0.3064	166.80
Hard OD gating	0.7140	0.2118	30.46	0.9338	0.3246	153.07
Continuous OD weighting (ours)	**0.7408**	**0.1994**	**29.35**	**0.9455**	**0.2391**	**145.15**

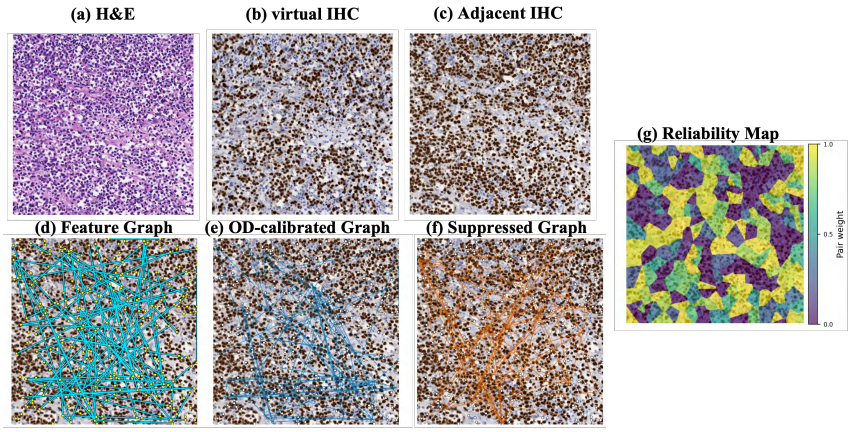

Figure 4: PECC graph calibration and candidate weighting. (a–c) H&E, virtual IHC, and adjacent real IHC. (d) Feature-affinity graph. (e,f) Feature edges retained and removed after OD calibration. (g) Cross-image OD-based candidate weights for $\mathcal{L}_{\text{corr}}$.

Continuous OD weighting achieves the best values on all six metrics within the fixed dual-topology design. Relative to uniform weighting, feature-similarity weighting improves HistCorr, leaves PearsonR unchanged at the reported precision, and worsens the remaining metrics. Hard gating improves FID and PathFID but worsens the other four metrics. These comparisons favor gradual OD-based attenuation over the tested feature-based and binary weighting rules, without establishing that OD agreement identifies biologically correct correspondences.

Figure 4 illustrates the complementary roles of graph calibration and candidate weighting. Panels (d–f) show how OD compatibility retains or suppresses feature-affinity edges within each image, thereby conditioning the context used to represent each sampled patch. Panel (g) shows how cross-image OD compatibility weights the supervision contributed by spatially proposed pairs. Thus, graph calibration modifies intra-image aggregation, whereas candidate weighting adjusts contrastive contributions without redefining the proposed correspondences.

A.1.6 Qualitative Module Ablation

Figure 5 complements the quantitative objective ablation in Table 3 with representative MIST–ER outputs. Following Pati et al. (2024), we include DeepLIIF-derived cell locations and predicted staining classes as auxiliary visualizations. These predictions are not cell-level reference annotations.

A.1.7 Implementation of Relational Matching Controls

All controls in Table 5 use the same node-construction procedure, two-dimensional OD-state descriptors, and normalized tissue-coverage masses. Let $\{(x_i, a_i)\}$ and $\{(y_j, b_j)\}$ denote the virtual and reference node sets, respectively. The shared Gaussian kernel is

$$k(x, y) = \exp\left[-\frac{1}{2}\,\|(x - y) \oslash \boldsymbol{\sigma}\|_2^2\right], \qquad \boldsymbol{\sigma} = (0.1, 0.25), \tag{30}$$

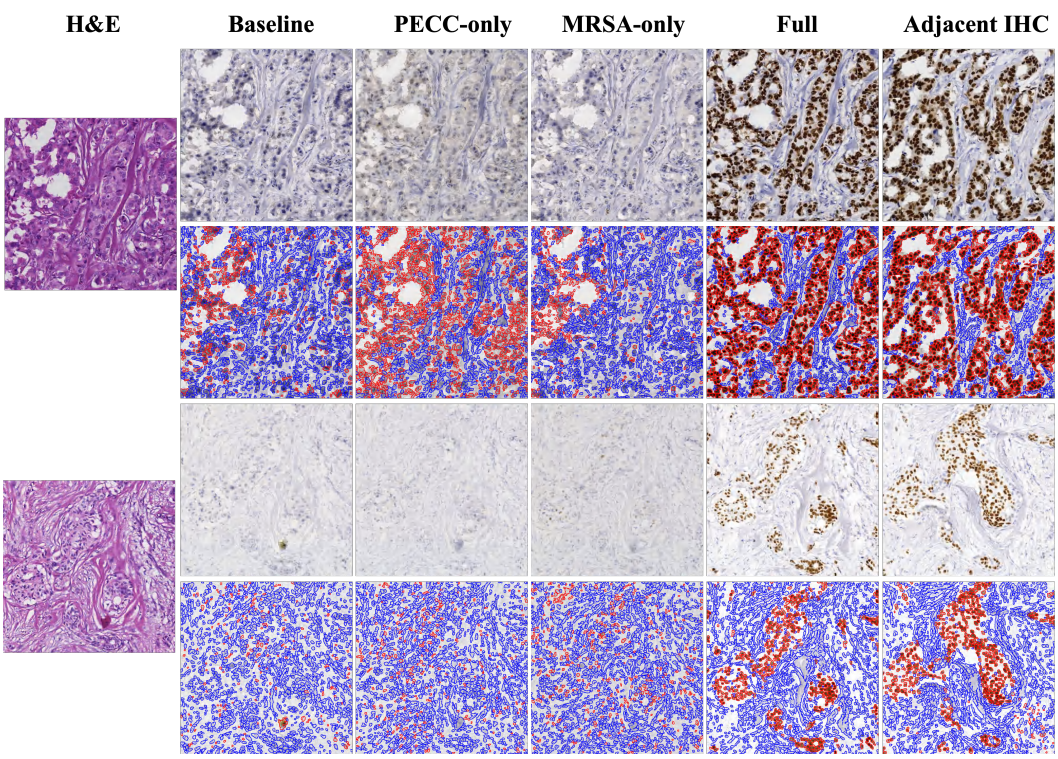

Figure 5: Qualitative module ablations on MIST–ER, accompanied by DeepLIIF-derived cell locations and staining-class predictions for auxiliary visualization.

where the bandwidths correspond to mean FOD and soft-positive pixel fraction.

Direct 2-D UOT. The cross-image cost is $D_{ij} = 1 - k(x_i, y_j)$. We estimate a coupling using the same generalized-KL marginal penalties and product-reference entropic regularization as UGW, with $\rho = 2$ and $\varepsilon = 0.05$. The direct transport problem uses 30 unbalanced Sinkhorn iterations, without GW linearization. The coupling is detached during generator optimization, which minimizes

$$\mathcal{L}_{\mathrm{UOT}} = \frac{\langle D, \boldsymbol{\pi}^* \rangle}{\|\boldsymbol{\pi}^*\|_1}. \tag{31}$$

Joint MMD. We use the biased, mass-weighted squared MMD estimator with the same kernel:

$$\begin{aligned} \mathcal{L}_{\mathrm{MMD}} = &\sum_{i,i'} a_i a_{i'} k(x_i, x_{i'}) + \sum_{j,j'} b_j b_{j'} k(y_j, y_{j'}) \\ &- 2 \sum_{i,j} a_i b_j k(x_i, y_j). \end{aligned} \tag{32}$$

The within-image sums include diagonal terms. No transport coupling is estimated, and small negative values caused by numerical error are clamped to zero.

Balanced and unbalanced GW. BGW and UGW use the same intra-image relation matrices, $\varepsilon = 0.05$, and six GW linearizations with 30 Sinkhorn iterations per linearization. BGW replaces the unbalanced scaling exponent $\rho/(\rho + \varepsilon)$ with one to target the prescribed marginals. Both use detached couplings and the squared-mass-normalized relational loss in Eq. 9. The fixed iteration budget provides approximate solutions.

Loss scaling. Each alternative replaces $\mathcal{L}_{\mathrm{rel}}$ with coefficient 1.0, while the remaining objectives and training settings are unchanged. Direct and relational transport losses are normalized by trans-

Table 8: Sensitivity of UGW to the marginal penalty ρ on MIST–ER, with $\varepsilon = 0.05$. The Full model without $\mathcal{L}_{\text{rel}}$ serves as the reference. PathFID denotes PathFID–StainNet. The best and second-best values are highlighted in **bold** and underlined, respectively.

Setting	HistCorr ↑	FID ↓	PearsonR ↑	DAB–KL ↓	PathFID ↓
Full w/o $\mathcal{L}_{\text{rel}}$	0.7235	31.29	0.9368	0.3103	170.83
UGW ($\rho = 0.10$)	<u>0.7253</u>	31.95	0.9281	0.3390	173.54
UGW ($\rho = 0.25$)	0.7223	30.84	0.9372	0.2973	161.29
UGW ($\rho = 0.50$)	0.7190	30.84	0.9410	0.2874	171.80
UGW ($\rho = 1.00$)	0.7236	<u>30.38</u>	<u>0.9415</u>	<u>0.2770</u>	<u>152.04</u>
UGW ($\rho = 2.00$)	**0.7408**	**29.35**	**0.9455**	**0.2391**	**145.15**
UGW ($\rho = 4.00$)	0.6985	31.35	0.9340	0.3971	163.09

Table 9: Sensitivity to the UGW regularization coefficient ε on MIST–ER with $\rho = 2$. All other settings are fixed. Full w/o $\mathcal{L}_{\text{rel}}$ is the non-relational reference. PathFID uses StainNet features. The best result in each column is highlighted in **bold**.

Setting	HistCorr ↑	DISTS ↓	FID ↓	PearsonR ↑	DAB–KL ↓	PathFID ↓
Full w/o $\mathcal{L}_{\text{rel}}$	0.7235	0.2185	31.29	0.9368	0.3103	170.83
UGW ($\varepsilon = 0.02$)	0.7232	0.2124	30.81	0.9389	0.3334	171.57
UGW ($\varepsilon = 0.05$, default)	**0.7408**	**0.1994**	**29.35**	**0.9455**	**0.2391**	**145.15**
UGW ($\varepsilon = 0.10$)	0.7176	0.2190	31.49	0.9390	0.3647	175.71
UGW ($\varepsilon = 0.20$)	0.7288	0.2098	30.51	0.9336	0.2847	164.60

ported mass and its square, respectively; MMD uses node masses normalized to sum to one. The comparison therefore shares descriptor scales and loss coefficients, but does not explicitly equalize loss magnitudes or gradient norms across formulations.

A.1.8 Sensitivity to Marginal Relaxation

We examine sensitivity to the marginal penalty ρ in the UGW coupling estimation used by $\mathcal{L}_{\text{rel}}$. The coupling is estimated through iterative GW linearization and entropic unbalanced Sinkhorn updates. The coefficient ρ penalizes deviations of the transported marginals from the prescribed relative node masses, while ε controls entropic regularization. Smaller ρ permits greater marginal deviations, whereas larger ρ penalizes these deviations more strongly.

We vary $\rho \in \{0.10, 0.25, 0.50, 1.00, 2.00, 4.00\}$ while fixing $\varepsilon = 0.05$ and all other model, optimization, and evaluation settings. The Full model without $\mathcal{L}_{\text{rel}}$ provides a common reference.

As shown in Table 8, $\rho = 0.10$ slightly improves HistCorr over removing $\mathcal{L}_{\text{rel}}$, but worsens FID, PearsonR, DAB–KL, and PathFID. A weak marginal penalty therefore does not provide a consistent improvement over the reference. The response to increasing ρ is non-monotonic: $\rho = 1.00$ improves all five metrics over the reference, and $\rho = 2.00$ achieves the best values across all tested settings, whereas increasing ρ further to 4.00 worsens all five metrics relative to $\rho = 2.00$.

These results support the default $\rho = 2.00$ within the evaluated range and show that stronger marginal constraints do not necessarily improve generation quality. Sensitivity to the entropic regularization coefficient ε is examined separately in the following subsection.

A.1.9 Sensitivity to Entropic Regularization

The coefficient ε controls the strength of coupling regularization toward the product reference measure in the UGW objective. Smaller values emphasize relational matching, whereas larger values encourage more diffuse couplings closer to the reference measure. We evaluate $\varepsilon \in \{0.02, 0.05, 0.10, 0.20\}$ in the Full model with $\rho = 2$, keeping all remaining model, training, and evaluation settings fixed, including graph construction and solver iteration budgets. Full without $\mathcal{L}_{\text{rel}}$ serves as the non-relational reference.

Table 9 shows that the default $\varepsilon = 0.05$ achieves the best results across all six metrics. Relative to the non-relational reference, $\varepsilon = 0.02$ improves DISTS, FID, and PearsonR but worsens HistCorr,

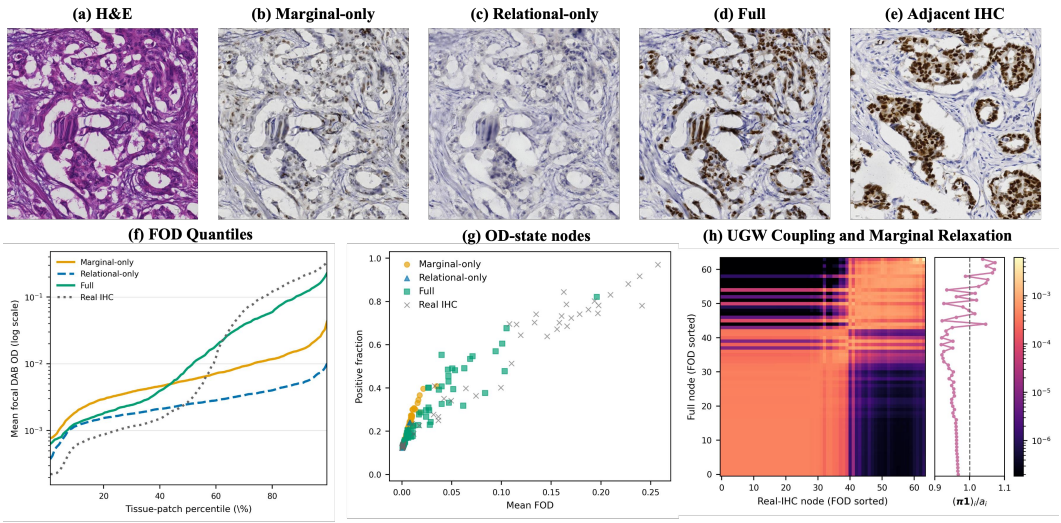

Figure 6: Visualization of MRSA on a weakly paired example. (a) H&E input. (b,c) PECC augmented with $\mathcal{L}_{\text{marg}}$ or $\mathcal{L}_{\text{rel}}$, respectively. (d) Full model using both objectives. (e) Adjacent real IHC. (f) Tissue FOD quantiles. (g) OD-state nodes $[\text{mean}(\text{FOD}), p^{+}]$, with node size indicating prescribed mass. (h) Estimated UGW coupling with nodes independently sorted by FOD, together with the source marginal ratio $(\boldsymbol{\pi}\mathbf{1})_i/a_i$. The dashed line at one indicates agreement with the prescribed source masses.

DAB–KL, and PathFID. Increasing ε to 0.10 worsens all metrics except PearsonR relative to the same reference. At $\varepsilon = 0.20$, five metrics improve over the reference, but PearsonR decreases, and all six remain worse than at the default setting. These non-monotonic responses indicate that the benefit of relational alignment depends on the regularization strength and support $\varepsilon = 0.05$ within the evaluated configuration.

A.1.10 Visualization of Marginal and Relational Alignment

Figure 6 illustrates the behavior of the two MRSA objectives on a weakly paired example. The adjacent real IHC provides a reference for staining composition and OD-state relations rather than a pixel-aligned target.

The marginal objective $\mathcal{L}_{\text{marg}}$ aligns the distribution of tissue-block mean FOD values, whereas $\mathcal{L}_{\text{rel}}$ compares pairwise relations among OD-state descriptors without fixed cross-section node assignments. In this example, combining both objectives produces closer visual agreement with the reference in the high-FOD quantiles (Figure 6(f)) and the OD-state distribution (Figure 6(g)), complementing the quantitative ablations.

Figure 6(h) displays the estimated coupling under relaxed marginal constraints. Nodes on the two axes are independently sorted by FOD for visualization; this ordering does not define cross-section correspondences. The accompanying profile compares each source coupling marginal $(\boldsymbol{\pi}\mathbf{1})_i$ with its prescribed mass a_i. A ratio of one indicates agreement with the prescribed mass, while values below or above one indicate less or more transported mass, respectively. The node-dependent departures from one visualize how the estimated coupling deviates from the prescribed source marginals.

A.1.11 Sensitivity to Loss Weights

We evaluate one-at-a-time sensitivity by halving or doubling each loss coefficient while keeping all remaining settings fixed. As shown in Table 10, the default configuration achieves the best HistCorr, DISTS, PearsonR, and PathFID across the tested settings. Doubling λ_{corr} slightly improves FID from 29.35 to 28.96, at the expense of the other four metrics. For λ_{struct}, λ_{marg}, and λ_{rel}, both tested deviations from the default degrade all reported metrics. These comparisons support the

default coefficients on MIST–ER while revealing metric-dependent sensitivity to the strength of correspondence supervision.

Table 10: Sensitivity to loss coefficients on MIST–ER. Within each group, only the indicated coefficient is varied; all remaining settings are fixed. $1\times$ denotes the default configuration, repeated for reference. The best result per metric within each group is highlighted in **bold**.

Coefficient	Scale (value)	HistCorr ↑	FID ↓	DISTS ↓	PearsonR ↑	PathFID–StainNet ↓
λ_{corr}	$0.5\times$ (0.05)	0.7323	30.49	0.2075	0.9438	152.46
	$1\times$ (0.10)	**0.7408**	29.35	**0.1994**	**0.9455**	**145.15**
	$2\times$ (0.20)	0.7365	**28.96**	0.2077	0.9442	149.29
$\lambda_{\mathrm{struct}}$	$0.5\times$ (0.05)	0.7328	29.75	0.2113	0.9384	158.44
	$1\times$ (0.10)	**0.7408**	**29.35**	**0.1994**	**0.9455**	**145.15**
	$2\times$ (0.20)	0.7378	31.04	0.2175	0.9429	156.54
λ_{marg}	$0.5\times$ (2.50)	0.7304	30.57	0.2160	0.9372	167.74
	$1\times$ (5.00)	**0.7408**	**29.35**	**0.1994**	**0.9455**	**145.15**
	$2\times$ (10.00)	0.7280	30.31	0.2032	0.9435	148.65
λ_{rel}	$0.5\times$ (0.50)	0.7242	29.99	0.2091	0.9446	156.26
	$1\times$ (1.00)	**0.7408**	**29.35**	**0.1994**	**0.9455**	**145.15**
	$2\times$ (2.00)	0.7354	30.54	0.2074	0.9450	159.92

A.1.12 Effectiveness of OD-State Structural Preservation

We ablate the degree, spectral, and three-step return terms in $\mathcal{L}_{\mathrm{struct}}$ while keeping the remaining objectives and training settings fixed. As shown in Table 11, the complete structural objective improves DISTS, FID, PearsonR, and PathFID over removing $\mathcal{L}_{\mathrm{struct}}$, with only a small change in HistCorr. The degree-only variant achieves the highest PearsonR but worsens the other four metrics relative to removing the structural objective. Adding the spectral term improves DISTS and PathFID over degree alone, while reducing HistCorr and PearsonR and increasing FID. Thus, progressively adding statistics does not yield uniform gains.

Adding the return term to the degree–spectrum objective improves all five metrics, yielding the best HistCorr, FID, DISTS, and PathFID among the tested configurations. Importantly, the return statistic satisfies $r^q = N^{-1}\sum_{j=1}^{N}(1-\xi_j^q)^3$ and is therefore determined by the full normalized Laplacian spectrum. Its contribution is an additional nonlinear spectral penalty, rather than structural information independent of the spectrum. The ablation supports this combined penalty formulation in the evaluated setting.

A.1.13 Computational Complexity

All proposed objectives are used only during training. The auxiliary graph and transport branches are discarded at inference, leaving the generator architecture and inference complexity unchanged.

For P sampled nodes with d-dimensional features, dense feature-affinity construction requires $\mathcal{O}(P^2d)$ operations and $\mathcal{O}(P^2)$ adjacency storage. Constructing a scalar OD-relation matrix requires $\mathcal{O}(P^2)$ operations. For $\mathcal{L}_{\mathrm{struct}}$, computing the full Laplacian spectrum and directly evaluating the three-step return statistic through dense matrix multiplication each require $\mathcal{O}(P^3)$ operations.

For UGW graphs with n and m nodes, each GW linearization requires $\mathcal{O}(n^2m+nm^2)$ operations, and each Sinkhorn iteration requires $\mathcal{O}(nm)$. We cap each graph at 64 nodes and use $T_{\mathrm{GW}}=6$ linearizations with $T_{\mathrm{SK}}=30$ Sinkhorn iterations per linearization. These caps bound the transport subproblem independently of image resolution; OD computation, tissue-mask construction, and block statistics still depend on the number of input pixels.

Algorithm 1 summarizes the finite-step coupling estimation. The relational cost compares independently constructed intra-image OD-state relations rather than direct cross-image feature distances. Coupling estimation is performed without gradient tracking, with denominator floors and clamped scaling ratios for numerical protection. During generator optimization, the estimated coupling and real-IHC relation matrix are detached, and gradients of the normalized relational loss propagate through the virtual-IHC relation matrix.

Table 11: Ablation of the structural terms in $\mathcal{L}_{\text{struct}}$ on MIST–ER. Deg, Spec, and Ret denote the degree, Laplacian-spectrum, and three-step return terms, respectively. None removes $\mathcal{L}_{\text{struct}}$. All remaining model components and training settings are fixed, with $\rho = 2$ and $\varepsilon = 0.05$. PathFID uses StainNet features. The best and second-best results per metric are highlighted in **bold** and underlined, respectively.

Structural statistics	HistCorr ↑	FID ↓	DISTS ↓	PearsonR ↑	PathFID ↓
None	<u>0.7400</u>	<u>30.05</u>	0.2102	0.9401	153.68
Deg	0.7359	30.23	0.2105	**0.9463**	154.27
Deg+Spec	0.7255	30.40	<u>0.2066</u>	0.9360	<u>153.12</u>
Deg+Spec+Ret	**0.7408**	**29.35**	**0.1994**	<u>0.9455</u>	**145.15**

A.1.14 Computational Resources and Runtime

We conduct the efficiency audit on MIST–ER using a single NVIDIA GeForce RTX 4090 GPU, 512×512 inputs, and batch size 1. Table 12 reports recorded per-epoch training times and deployment complexity. The Full model requires 23.88 minutes per epoch, compared with 15.37 minutes for the backbone baseline, corresponding to approximately $1.55\times$ the training time.

At inference, all configurations use the same six-block ResNet generator without the auxiliary loss branches. We measure generator complexity using `fvcore` with a $1 \times 3 \times 512 \times 512$ input tensor, counting one fused multiply–add as one floating-point operation. Under this convention, all configurations require 199.10 GFLOPs and contain 7.838 million inference parameters. The proposed objectives therefore increase training cost without adding inference operations or parameters.

Table 12: Computational-resource audit on MIST–ER at 512×512 resolution with batch size 1. Relative training time is normalized by the backbone baseline. Inference FLOPs and parameter counts refer only to the deployed generator; auxiliary branches are discarded after training.

Method	Training time (min/epoch) ↓	Relative training time	Inference GFLOPs ↓	Inference parameters (M) ↓
Baseline	15.37	1.00×	199.10	7.838
PECC	20.65	1.34×	199.10	7.838
MRSA	18.65	1.21×	199.10	7.838
Full (Ours)	23.88	1.55×	199.10	7.838

A.2 Additional Quantitative and Qualitative Results

The main paper compares PGLStain with three general image-translation methods CycleGAN Zhu et al. (2017), CUT Park et al. (2020), and UNSB Kim et al. (2024) and eight virtual staining methods: PyramidP2P Liu et al. (2022), ASP Li et al. (2023), PPT Zhang et al. (2024), SIMGAN Guan et al. (2025), PSPStain Chen et al. (2024), MDCL Wang et al. (2024), M2PLGAN Liu et al. (2026), and USIGAN Peng et al. (2026). Here, we report additional results for MIST HER2, MIST PR, MIST Ki67. Figures 8– 10 provide the corresponding qualitative comparisons.

Across the three additional MIST stains, PGLStain provides consistent gains in paired perceptual quality and OD-related agreement. It obtains the best PHV, DISTS, and PearsonR for all three stains; it also achieves the best HistCorr on HER2 and PR and the best DAB KL on HER2 and Ki67. On metrics for which another method is best, PGLStain remains competitive, including the second-best PathFID on HER2 and Ki67 and the second-best FID on Ki67. These results indicate that the improvement observed on MIST–ER extends across distinct biomarker-expression patterns, although the optimal result is not uniform across every metric.

A.3 Loss Responses to DAB-Tile Permutation

A.3.1 Purpose and Construction

We examine how the frozen model's objectives respond to spatial rearrangement of DAB staining while preserving its decomposed value distribution. The analysis uses all 1,000 patches in the MIST–

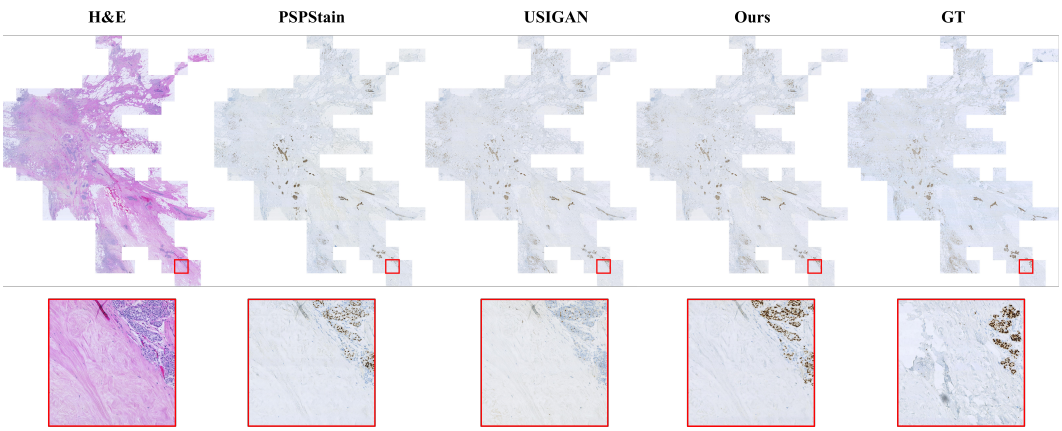

Figure 7: WSI-level qualitative comparison on ACROBAT–ER dataset.

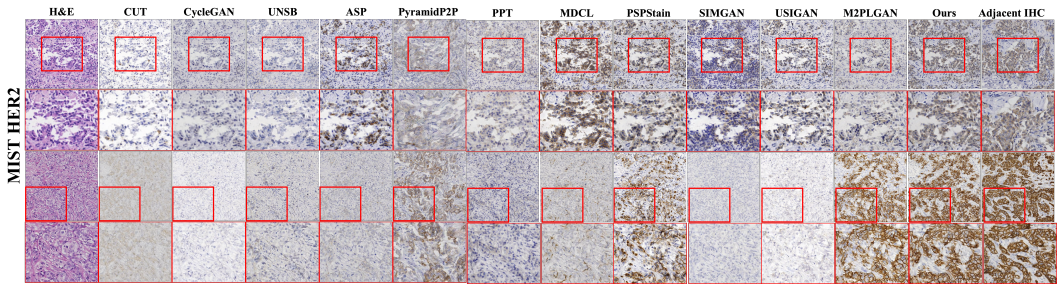

Figure 8: Qualitative comparison on MIST–HER2. All methods are evaluated on the same input patches and displayed using the same resolution.

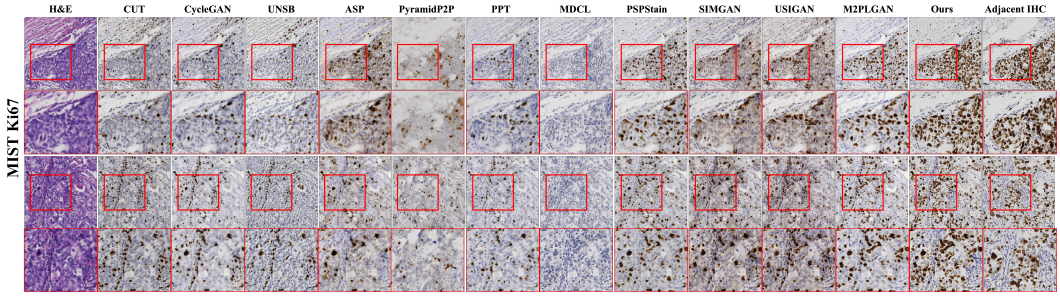

Figure 9: Qualitative comparison on MIST–Ki67. All methods are evaluated on the same input patches and displayed using the same resolution.

ER validation set. Each patch is treated as an input image for perturbation and loss evaluation, with its weakly paired real-IHC reference held fixed. The experiment characterizes objective responses rather than testing invariance to changes in image content.

Given a generated IHC patch $\hat{Y}$, we perform HED deconvolution and recompose the image without modification to obtain a round-trip control $\hat{Y}^{\mathrm{ctrl}}$. For the counterfactual $\hat{Y}^{\mathrm{cf}}$, the H/E channels remain fixed, while the DAB channel is partitioned into non-overlapping tiles and permuted on 2×2, 4×4, or 8×8 grids. This operation preserves the DAB-value multiset in the decomposed representation while changing its placement relative to the fixed H/E components. RGB recomposition and subsequent deconvolution may introduce numerical deviations.

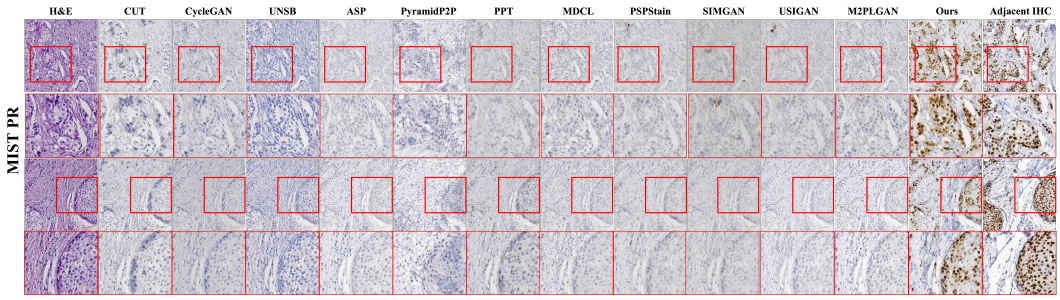

Figure 10: Qualitative comparison on MIST–PR. All methods are evaluated on the same input patches and displayed using the same resolution.

Table 13: Quantitative comparisons on the remaining MIST biomarkers. The best and second-best available results for each stain are highlighted in **bold** and underlined, respectively.

Dataset	Method	HistCorr ↑	$\mathrm{PHV}_{\mathrm{Avg}}$ ↓	DISTS ↓	FID ↓	PearsonR ↑	$\lvert\mathrm{Bias}_{\mathrm{mIoD}}\rvert$ ($\times 10^3$) ↓	DAB–KL ↓	PathFID–StainNet ↓
$\mathrm{MIST}_{\mathrm{HER2}}$	CUT	0.6257	0.5978	0.2528	37.24	0.4569	4.745	1.2533	233.08
	CycleGAN	0.6507	0.5837	0.2510	50.53	0.2680	32.961	0.9527	480.95
	UNSB	0.5723	0.6266	0.2598	37.30	0.4324	**4.077**	1.4447	257.29
	PyramidP2P	0.6889	0.5683	0.2515	62.12	0.8508	20.792	0.3363	434.60
	ASP	0.6591	0.5745	0.2504	46.90	0.6875	22.485	0.6478	429.50
	PPT	0.7024	0.5465	0.2360	44.29	0.6972	4.681	0.5783	338.46
	PSPStain	0.7309	0.5232	0.2099	**35.11**	0.8886	6.103	0.3728	**160.08**
	MDCL	0.6619	0.5256	0.2096	39.71	0.8537	12.130	0.3495	200.10
	SIMGAN	0.6363	0.6083	0.2451	39.35	0.7704	7.140	0.9758	209.49
	USIGAN	0.6993	0.5477	0.2204	38.11	0.8112	7.627	0.7671	189.78
	M2PLGAN	0.7361	0.5475	0.2201	37.13	0.8461	6.861	0.7288	180.82
	PGLStain (Ours)	**0.7521**	**0.5160**	**0.2054**	37.17	**0.8979**	6.197	**0.2961**	166.91
$\mathrm{MIST}_{\mathrm{PR}}$	CUT	0.5873	0.6273	0.2728	89.40	0.2213	58.238	0.7952	1480.55
	CycleGAN	0.6584	0.5923	0.2629	46.81	0.4594	35.146	1.2042	386.77
	UNSB	0.5642	0.6513	0.2758	37.85	0.5852	11.427	1.6216	547.29
	PyramidP2P	0.7201	0.5694	0.2477	50.68	0.9015	18.614	**0.2819**	374.28
	ASP	0.6697	0.5583	0.2350	34.29	0.8167	23.971	0.6566	234.04
	PPT	0.6777	0.5435	0.2364	35.40	0.8128	9.604	0.6472	307.18
	PSPStain	0.7232	0.5275	0.2092	29.83	0.9271	6.953	0.3593	155.15
	MDCL	0.7121	0.5281	0.2140	31.54	0.9119	10.369	0.3597	185.42
	SIMGAN	0.7101	0.5605	0.2262	30.73	0.8953	4.726	0.7500	160.53
	USIGAN	0.7012	0.5618	0.2209	30.70	0.8880	5.310	0.6144	163.14
	M2PLGAN	0.7446	0.5485	0.2124	**29.76**	0.9197	5.987	0.5606	150.16
	PGLStain (Ours)	**0.7533**	**0.5261**	**0.2070**	30.42	**0.9404**	**4.199**	0.2923	**148.15**
$\mathrm{MIST}_{\mathrm{Ki67}}$	CUT	0.6465	0.5775	0.2365	26.62	0.6446	3.281	0.7175	198.35
	CycleGAN	0.6746	0.5751	0.2369	34.95	0.4935	3.550	0.8185	283.33
	UNSB	0.4757	0.6353	0.2585	36.27	0.3916	4.229	1.0375	426.57
	PyramidP2P	0.6445	0.5722	0.2553	64.60	0.8948	12.734	0.2597	385.56
	ASP	0.6372	0.5574	0.2247	30.89	0.8209	5.279	0.4510	261.80
	PPT	0.6970	0.5367	0.2230	27.79	0.7518	3.517	0.6025	192.72
	PSPStain	0.7238	0.5225	0.2105	25.40	0.8987	2.260	0.2841	**133.94**
	MDCL	**0.7309**	0.5219	0.2088	25.86	0.8880	0.591	0.2752	158.36
	SIMGAN	0.6398	0.5741	0.2338	26.60	0.8084	3.013	0.6235	176.47
	USIGAN	0.6680	0.5525	0.2219	**24.63**	0.8430	**0.280**	0.5961	148.19
	M2PLGAN	0.7091	0.5448	0.2203	26.97	0.8272	4.616	0.5917	167.72
	PGLStain (Ours)	0.7285	**0.5154**	**0.2051**	25.05	**0.9321**	0.526	**0.2429**	141.52

Within each control–counterfactual comparison, we use the same image pair, PatchNCE sampling indices, and random sampling indices for $\mathcal{L}_{\mathrm{struct}}$. Features, OD values, and the tissue-filtered block descriptors used by MRSA are recomputed for each condition. No model parameter is updated. OD-aware and uniform correspondence weighting are evaluated using the same frozen representations within each condition, rather than separately trained models. We report output-level responses for the 2×2 perturbation and descriptive loss-response distributions at all three scales.

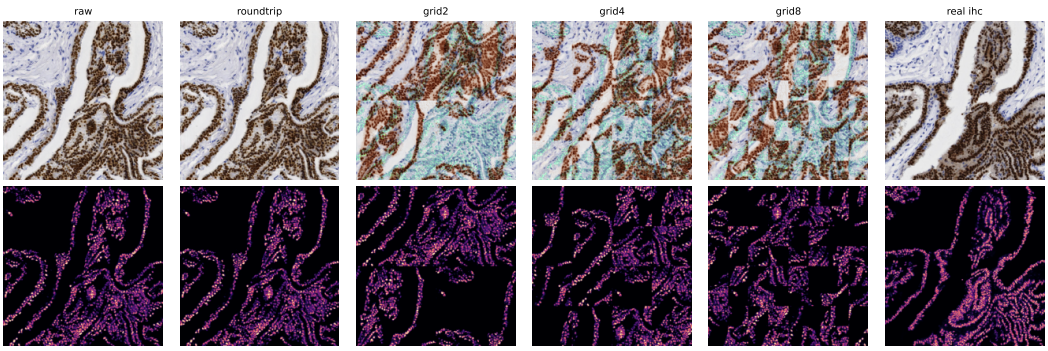

Figure 11: DAB-tile perturbations of MIST–ER patch 47M2101316_15_16. The figure shows the real-IHC reference, original virtual IHC, HED round-trip control, and 2×2, 4×4, and 8×8 DAB permutations with their FOD maps. The perturbations preserve the decomposed DAB-value multiset while rearranging staining relative to the fixed H/E components.

Table 14: Output-level response to DAB-tile permutation on 1,000 MIST–ER patches. The control uses HED deconvolution and recomposition without permutation; the counterfactual permutes DAB tiles on a 2×2 grid while retaining the H/E channels. Both conditions use the same reconstruction pipeline. The absolute values are not directly comparable with the raw-image results in the main paper. Relative changes are computed before rounding.

Condition	$\mathrm{PHV}_{\mathrm{Avg}} \downarrow$	DISTS↓	FID↓	PearsonR↑	$\lvert\mathrm{Bias}_{\mathrm{mIoD}}\rvert$ $(\times 10^3) \downarrow$	DAB KL↓	PathFID–StainNet↓
HED round-trip control	0.5451	0.2150	32.24	0.945216	13.332	0.2425	715.73
DAB shuffle (2×2)	0.6002	0.2614	60.50	0.945229	13.343	0.2428	1175.75
Relative change	+10.1%	+21.6%	+87.7%	+0.001%	+0.08%	+0.11%	+64.3%

A.3.2 Qualitative Perturbation Examples

Figure 11 illustrates the perturbation at different tile scales. Finer grids rearrange smaller DAB regions without introducing new values into the decomposed DAB channel. The example illustrates the construction; all reported evaluation-set statistics and loss distributions use the complete 1,000-patch validation set.

A.3.3 Conservation and Output-Level Response

The conservation audit verifies that the sorted DAB values and H/E channels are unchanged in the decomposed representation. After RGB recomposition, the mean absolute discrepancy between sorted FOD values is at most 2.75×10^{-5} across the evaluated patches and perturbation scales. These checks concern global stain-value distributions within each patch; they do not establish preservation of the sampled descriptors, tissue masks, or node sets used by the training objectives.

Table 14 compares the round-trip control with the 2×2 perturbation over the same evaluation set. PearsonR, DAB KL, and the absolute mean mIoD bias change by at most 0.11%, indicating little change in these evaluated staining statistics. PHV, DISTS, FID, and PathFID–StainNet increase, indicating reduced agreement with the unchanged real-IHC reference. Because HED reconstruction is not an identity operation, these comparisons use the round-trip control rather than the original generated images.

A.3.4 Response of the Training Objectives

For an objective quantity $\mathcal{Q}$, the paired response of patch i is

$$\Delta\mathcal{Q}_i = \mathcal{Q}(\hat{Y}_i^{\mathrm{cf}}, Y_i) - \mathcal{Q}(\hat{Y}_i^{\mathrm{ctrl}}, Y_i), \tag{33}$$

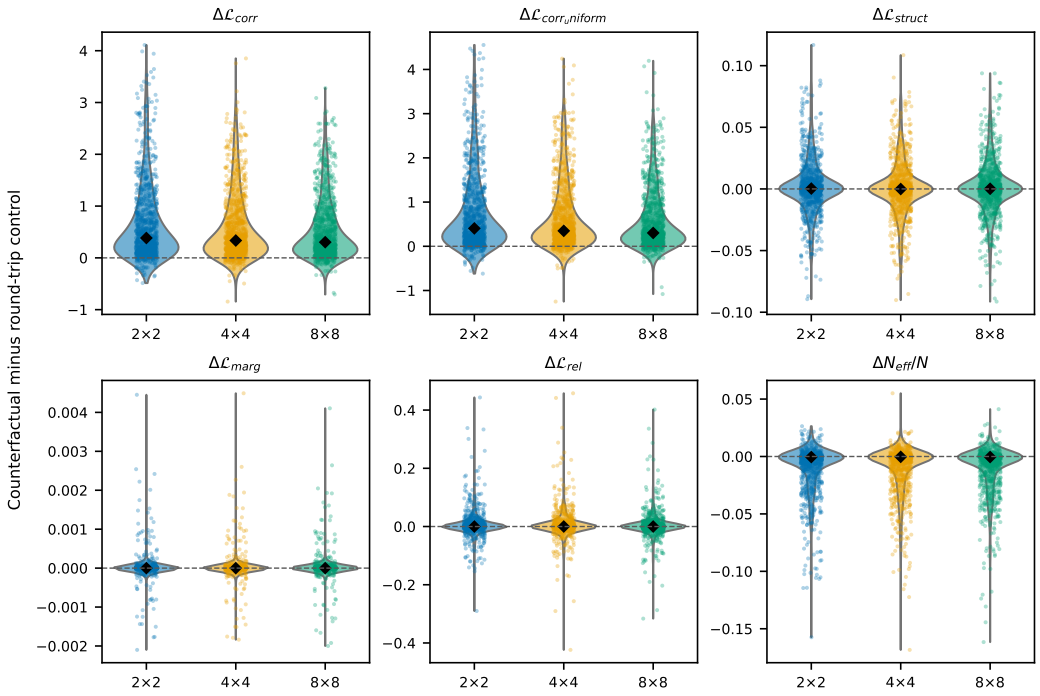

Figure 12: Patch-level distributions of counterfactual-minus-control responses under 2×2, 4×4, and 8×8 DAB permutations on the 1,000-patch MIST–ER validation set. The dashed line denotes zero. The distributions provide descriptive response diagnostics; response magnitudes should not be compared directly across objectives with different scales.

where Y_i is the unchanged weakly paired real-IHC reference. We also evaluate the weighted auxiliary objective

$$\mathcal{L}_{\text{aux}} = 0.1\mathcal{L}_{\text{corr}} + 0.1\mathcal{L}_{\text{struct}} + 5\mathcal{L}_{\text{marg}} + \mathcal{L}_{\text{rel}}. \tag{34}$$

This quantity excludes the GAN, PatchNCE, and Gaussian-pyramid terms and is not the complete generator objective.

Figure 12 summarizes the variation in paired responses across input patches at each perturbation scale. Positive and negative values indicate increases and decreases relative to the round-trip control, respectively. Comparing OD-aware and uniform correspondence weighting isolates how the weighting rule changes the response of the frozen model; it does not identify which candidate pairs are correctly matched or distinguish reference mismatch from prediction error. The loss distributions are reported descriptively, without treating the patches as independent inferential replicates. Small responses do not establish equivalence or invariance to the perturbation.

For a set of N nonnegative candidate weights w_j with $\sum_j w_j > 0$, weight concentration is characterized by the normalized effective sample size

$$\frac{N_{\text{eff}}}{N} = \frac{(\sum_j w_j)^2}{N \sum_j w_j^2}. \tag{35}$$

This quantity equals one for uniform nonzero weights and decreases as weights become more concentrated. It describes the distribution of supervision weights, not the fraction of correctly matched or retained candidates.

Node-permutation invariance concerns consistent reindexing of a fixed set of descriptors, masses, and relations. DAB-tile permutation instead changes the image content at sampled locations and can alter tissue-filtered block descriptors and node selection. The experiment therefore measures responses to within-patch stain rearrangement rather than testing node-reindexing invariance.